Apprentissage d'appariements pour la discrimination de séries temporelles

Cédric Frambourg

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Modèles, méthodes et algorithmes en biologie, santé et environnement (MBS)**

Arrêté ministérial :

Présentée par

Cédric Frambourg

Thèse dirigée par **Jacques Demongeot**
et codirigée par **Ahlame Douzal**

préparée au sein **TIMC-IMAG (Techniques de l'Ingénierie Médicale et de la Complexité - Informatique, Mathématiques et Applications de Grenoble)**
et de **Ecole Doctorale Ingénierie pour la Santé, la Cognition et l'Environnement.**

Apprentissage d'appariements pour la discrimination de séries temporelles

Thèse soutenue publiquement le **13 mars 2013**,
devant le jury composé de :

Antoine Cornuéjols
Professeur des Universités, AgroParisTech, Rapporteur
Eric Gaussier
Professeur des Universités, Université J. Fourier Grenoble, Président
Mohamed Nadif
Professeur des Universités, Université Paris Descartes, Examinateur
Fabrice Rossi
Professeur des Universités, Université Paris 1 Panthéon-Sorbonne, Examinateur
Marc Sebban
Professeur des Universités, Université Hubert Curien, Rapporteur
Jacques Demongeot
PU-PH, Université J. Fourier Grenoble, CHU Grenoble, Directeur de thèse
Ahlame Douzal
Maître de conférence HDR, Université J. Fourier Grenoble, Co-Directeur de thèse

Quand vient le moment des remerciements, la première préoccupation est de n'oublier personne. Je m'en sortirai par une pirouette en remerciant en avant-propos, tous ceux qui, par mégarde, ne figureraient pas dans la suite de manière nominative.

Je remercie tout d'abord profondémment Ahlame, qui m'a encadré pendant ces nombreuses années de thèse, et même avant, dès le master, qui m'a guidé tout le long du chemin, avec beaucoup d'énergie et surtout beaucoup de patience et de persévérance. Son investissement dans cette thèse a souvent été mon principal moteur dans mes moments de doutes. J'ai découvertr durant cette thèse un univers que je ne connaissais pas, et quand mes collègues viennent me voir en tant que statisticien, ce n'est que grâce à toi. Merci Jacques d'avoir suivi mes travaux, nos discussions ont toujours été très agréables, merci pour tes conseils et tes remarques.

Merci Eric pour ton investissement dans ces travaux, travailler avec toi était toujours très constructif, tu m'as encouragé à plus de formalisme, tu as redonné du sens à mes travaux à un moment où je me dispersai. Merci également d'avoir accepté de participer à mon jury ; sans toi, ce manuscrit n'aurait pas cette allure.

Je remercie également tous les autres membres du jury, Antoine Cornuéjols et Marc Sebban, je mesure la chance que j'ai eu de vous avoir pour rapporteurs. Merci pour vos remarques sur le manuscrit mais également durant la soutenance, et pour l'intérêt que vous portez à ces travaux. Merci également à Fabrice Rossi et Mohamed Nadif d'avoir accepté de faire partie de mon jury.

J'ai eu l'occasion de rejoindre durant cette thèse une nouvelle équipe. J'ai remercié Eric pour son encadrement et pour sa participation au jury. J'aimerais maintenant remercier Eric, le chef de l'équipe AMA, Père Castor si je reprend un surnom qui n'aura duré que deux jours. Participer à cette expérience avec tous les collègues devenus amis a été un grand moment dans ma thèse. Koko, ta brique pleine d'entrain va me manquer, ainsi que les nombreuses (trop diraient certains) pauses qui ponctuaient nos journées. Je continuerai à acheter Charlie le jeudi, jour de sa sortie. Koala, même si tu restes un petit jeunôt pour nous, je suis content de t'avoir accompagné dans l'adolescence. Et si je devais te dire quelque chose, c'est qu'il est important, ..., bref. Grincheux, si je devais te remercier pour une seule chose, c'est d'avoir réussi à faire renaître Bourg, mais il y a plein d'autres choses, merci pour nos virées touristiques, pour nos discussions politiques, pour tes conseils avisés (en particulier en ce qui concerne la grammaire), pour nos parties de coinches. Curi, tu as redéfini pour moi le terme glander au boulot, mais c'était un plaisir de travailler avec toi. Je ne verrai plus jamais un plat de poulet de la même façon. Joe, même si tu parles beaucoup et très vite, demerdes-toi pour venir me rejoindre l'an prochain, on va leur montrer la vie à ces bretons. Même si je dois pour cela rester le seul à me cogner aux pancartes trop basses. Alvy, ta danse à l'heure du repas mettait un peu de piquant dans la période de rédaction. Merci Cécile pour ton humour, j'ai adoré partager l'enseignement avec toi. Cornélia, Parantapa, Rohit, vous avez échappé aux surnoms idiots, merci pour votre bonne humeur. Et Parantapa, veille sur le sachet de thé caramel, en lui réside l'esprit amajunior.

Je n'oublie pas mes anciens collègues de Taillefer. Hédi, collègue, puis colloc, mais toujours ami. Nos moments passés à se plaindre à l'appart sur nos thèses me manquent déjà. Flora et Laure qui partagez avec moi l'expérience des meurtrières en guise de fenêtre. Nicolas, qui n'a jamais remarqué que l'aiguille des heures faisait deux fois le tour du cadran chaque journée. Merci Céline, de n'avoir jamais osé me foutre à la porte de ton bureau, même quand tu voulais

travailler. Merci à toi Adeline, de m'avoir soutenu à chaque fois que j'avais des messages de la part du rectorat. Et gardes toujours en mémoire que Superman est plus fort que Wonder Woman.

A tout mes amis de l'Institut Fourier, vous avez supporté un documentaire tournant en boucle sur l'Alsace pendant trois ans. Je vous en félicite et m'en excuse. Caroline, désolé pour tous les cours assis à côté de toi. Max, Camille, même si vous n'avez toujours pas remarqué que j'ai des poils sur le menton, intrinsèquement, c'était vraiment chouette de partager ces années avec vous. Marianne, tu as été ma première amie à Grenoble, tu l'es restée depuis. A cause de toi, j'ai été triste de quitter la ville. Qui l'aurait cru 8 ans plus tôt. Merci à Aude, pour nos discussions entre deux avions. Thomas, danke für deine Freundschaft. Mein Deutsch ist viel besser heute.

Je remercie également tous mes collègues Vannetais de l'IUT, et mes collègues de l'an dernier à Grenoble. Je n'ai pas dû être facile à vivre à certaines périodes de ces deux années. Merci de m'avoir épaulés et poussés à aller jusqu'au boût. Merci pour votre soutien pour les cours. J'ai trouvé grâce à vous un domaine qui me plait.

Je remercie enfin tous mes amis et ma famille de m'avoir offert un cadre idéal pour mener à bien mes travaux.

Et toi, qui n'a pas voulu te reconnaitre dans mon avant-propos, je souhaite te remercier, ..., pour avoir contribué à un moment ou à un autre au bon déroulement de cette thèse.

Table des matières

Table des figures

Introduction

Motivations

Les séries temporelles sont présentes dans de nombreux domaines d'application. En particulier, le problème de la classification des séries temporelles est une tâche importante. Beaucoup de méthodes usuelles pour la classification nécessitent de pouvoir comparer les séries. La plupart de ces procédés de comparaisons sont dérivées de la DTW (Dynamic Time Warping) et alignent les observations des paires de séries de manière monotone pour repérer les délais qui peuvent apparaître entre les instants. La comparaison de séries temporelles repose en général sur les deux principes suivants : une comparaison paire à paire de tous les couples de séries et un appariement global de toutes les observations de chaque série. Les deux principes supposent implicitement que les séries temporelles partagent un profil très proche au sein des classes.

Dans les applications réelles, les séries temporelles peuvent présenter des caractéristiques beaucoup plus complexes. Il n'est pas rare que les séries d'une même classe aient des profils très différents. Au contraire, il n'est pas rare également pour des séries de classes différentes d'avoir des profils similaires. Les séries d'une classe peuvent partager une signature discriminante locale, tel un évènement saillant apparaissant à un instant qui fluctue. Cette signature caractérise les séries d'une classe et les différencie des séries des autres classes. Une telle complexité au sein des classes de séries rend limitées les approches usuelles. Il est alors crucial que le couplage se fasse en adéquation avec le processus de discrimination. Pour discriminer des séries complexes, il est important de s'intéresser à la dynamique de l'ensemble des séries au sein et entre les classes. Il est également fondamental de privilégier les instants les plus discriminants au cœur de l'analyse.

Des réponses à ces questions ont été amorcées dans plusieurs travaux. Le problème du couplage paire à paire est assez ancien et demeure toujours un défi. Pour l'alignement de séquences peptidiques associées à des protéines partageant une même fonction, les biologistes cherchent à aligner les séquences selon une dynamique commune à l'ensemble. Les solutions proposées dans ce contexte consistent généralement à rassembler les alignements paires à paires obtenus pour étendre le processus à un alignement des multiples séquences. Afin de lier l'appariement des observations à la tâche de discrimination, certains travaux récents s'affranchissent des alignements de la DTW. Ramsay et Li (1998) proposent de définir une fonction de délai pour chaque série, pour aligner les courbes en cherchant à préserver la courbure de la série initiale. Dans le même esprit, Gaffney et Smyth (2005) proposent un apprentissage probabiliste des appariements fondé sur une approche EM. Cependant, les deux approches se limitent à la comparaison de séries de la même classe. Listgarten *et al.* (2007) recherchent les éléments distinctifs entre les différentes classes de séries. Leur approche

consiste en un modèle bayésien hiérarchique fondé sur des modèles de Markov cachés. Toutes ces approches échouent cependant à discriminer les classes lorsque les séries au sein d'une classe présentent un profil général différent.

Le critère de la variance/covariance locale (Lebart, 1969) est très utilisé pour des tâches de discrimination. On le retrouve notamment dans le cadre de l'analyse de la contiguïté (Geary, 1954; Moran, 1950; Wartenberg, 1985; Thioulouse *et al.*, 1995). Mom (1988) étend la notion de variance locale à une partition, dans le cadre de l'étude d'un réseau de transport.

Contribution de cette thèse

Dans le cadre de cette thèse, nous menons dans un premier temps une étude bibliographique des mesures de proximité entre séries temporelles. Nous nous intéressons à l'analyse de données contiguës, dont les séries temporelles sont un cas particulier. A travers l'étude des indices d'autocorrélation spatiale de Moran (1950) et Geary (1954), nous présentons le lien existant entre la structure du voisinage et les observations. La variance locale est un indicateur calculée sur l'ensemble des observations et n'est pas fondée sur une approche paire à paire. Nous étendons la variance locale à travers un nouveau formalisme, à un ensemble, puis à une partition de séries temporelles. Chaque appariement entre les instants des séries temporelles induit une valeur intrinsèque de la variance. Nous étudions alors la contribution d'une arête à la variance.

Dans un second temps, cette contribution à la variance est utilisée en vue de l'apprentissage d'appariements discriminants. Nous proposons une méthode d'apprentissage d'appariements optimisant le problème de minimisation de la variance intra et de maximisation de la variance inter. Nous introduisons deux familles de contraintes qui rendent les deux problèmes d'optimisation tantôt convexes, tantôt discrets. L'approche proposée consiste à pénaliser de manière itérative les arêtes liant deux instants en fonction de leur contribution à la variance. A partir d'un appariement initial donné, les arêtes sont sélectionnées puis pénalisées jusqu'à l'optimalité. Nous démontrons que notre approche est équivalente d'un point de vue calculatoire à la méthode du gradient projeté, qui est la méthode usuelle pour ce type de problème. L'apprentissage, au contraire des approches usuelles, est fait en tenant compte de la dynamique de l'ensemble des séries, privilégiant les instants localement discriminant, et est effectué au regard du problème de discrimination considéré. Nous proposons finalement, pour l'apprentissage d'appariements discriminants, une méthode d'enchevêtrement des processus intra et inter classes. Nous définissons, à partir des appariements discriminants appris, une métrique associée à un ensemble de poids discriminants. Cette métrique montre la pertinence des appariements appris en vue de la discrimination de séries temporelles complexes.

Organisation du manuscrit

Le présent manuscrit se déroule de la manière suivante.

La première partie étudie les méthodes existantes pour la classification et la discrimination de séries temporelle. Le premier chapitre présente les approches classiques de classification. La plupart de ces méthodes reposent sur la comparaison paire à paire des séries temporelles à partir de l'alignement de l'ensemble de leurs observations. Dans le second chapitre, nous étudions la variance locale, fréquemment utilisée pour des problèmes de discrimination, et

son extension à un ensemble puis à une partition de séries temporelles. Ce chapitre note l'importance des appariements dans la définition de la variance locale.

La seconde partie propose une méthode d'apprentissage des appariements en vue de minimiser la variance intra et de maximiser la variance inter classe. Le premier chapitre formalise les deux problèmes d'optimisation. Il introduit deux types de formalisme ramenant le problème d'optimisation tantôt à un problème convexe, tantôt à un problème discret. Le seconde chapitre détaille trois algorithmes. Le premier vise à apprendre des appariements minimisant la variance intra ; le second vise à maximiser la variance inter, et un troisième algorithme permet de rassembler les deux processus intra et inter en vue de l'apprentissage d'appariements discriminants. Les aspects calculatoires liés à ces trois algorithmes sont également présentés.

La troisième partie montre sur des jeux de données complexes la pertinence des appariements appris. Le premier chapitre définit des poids liés à l'entropie des appariements appris puis introduit une métrique discriminante fondée sur ces poids et sur les appariements. Cette métrique est étudiée sur la base d'une classification KPPV (k plus proches voisins) pour des jeux de données simulés. Le second chapitre étudie les résultats de cette métrique pour un jeu de données réelles, décrivant la consommation journalière d'un foyer tout au long d'une année, selon deux problématiques, la prédiction précoce d'un pic de consommation, et l'extraction d'une tendance saisonnière.

Partie I

Positionnement des travaux

La première partie de ce manuscrit situe le contexte de ce travail. Notre objectif est la comparaison et la classification de séries temporelles. Le premier chapitre présente les séries temporelles et s'intéresse aux méthodes usuelles employées pour comparer des séquences ou des séries temporelles en vue de leur classification. Nous remarquons que de nombreuses méthodes reposent sur l'alignement des instants des séries comparées. Pour traiter des problèmes de classification, les matrices de variance-covariance intra et inter-classes sont classiquement utilisées. Nous décrivons dans le second chapitre les travaux menés sur les données contiguës visant à étendre les matrices de variance-covariance intra et inter-classes à des données structurées. Nous proposons finalement une extension de la variance au cas d'ensembles et de partitions de séries temporelles. Nous montrons alors l'importance qu'occupe le choix des appariements.

Chapitre 1

Comparaison et classification de séries temporelles

Les séquences temporelles et les séries temporelles sont deux objets fréquemment rencontrés. Les séries temporelles pouvant être vues comme un cas particuler de séquences, nos recherches sur les séries temporelles nous ont amenés à nous intéresser à la notion de séquence. Nous faisons dans un premier temps un état de l'art des mesures de comparaison entre paires de séquences. Dans un second temps, nous présentons sous un formalisme unifié les méthodes usuelles de comparaison de séries temporelles. Une mesure de similarité repose en général sur trois définitions : une distance entre les observations, un alignement des observations et une fonction de coût associée à cet alignement. Cependant, l'alignement des paires de série ne reflète pas les propriétés des classes. Nous évoquons alors plusieurs travaux visant à l'alignement simultané d'un ensemble de séries temporelles afin de rechercher les structures communes au sein des classes. Nous étudions dans une dernière partie comment les mesures de proximité sont mises en application pour la classification de séries temporelles.

Nous désignons par données temporelles des données numériques évoluant dans le temps, dites communément séries temporelles, ou des suites chronologiques de données symboliques dites séquences temporelles. Plus généralement, on désigne par données de séquences toute collection de données ordonnées selon un critère qui peut être sémantique, biologique, temporel ou autre ; c'est le cas, par exemple, des séquences de mots dans un texte ; on parle alors d'ordre syntaxique, de séquences d'acides aminés composant une chaîne d'ADN ou de peptides constituant une protéine.

1 Comparaison de séquences

Les séries et les séquences temporelles ont une structure similaire. Dans le cadre de ce travail de thèse sur les séries temporelles, nous nous sommes dans un premier temps intéressés aux travaux existants dans le contexte des séquences temporelles. Ces travaux visent à comparer deux séquences. Pour cette comparaison, la plupart des approches étudiées font

appel à des mesures de proximité. Nous débuterons cette section par la définition de la notion générale de mesure de proximité. La plupart des mesures de proximité usuelles entre séquences sont fondées sur la notion d'alignement. Nous introduisons dans la suite de cette section quelques notations quant aux alignements et quelques mesures usuelles de proximité entre séquences, fondées sur ces alignements.

1.1 Mesures de proximité entre séquences

Généralités Notons Ω un ensemble, et ω son élément courant ; dans notre cas, Ω est l'ensemble des séquences. Les mesures de proximité sont des applications qui, à chaque couple (ω_1, ω_2) de Ω^2, associent une quantité mesurant, comme son nom l'indique, leur proximité. Nous présentons tout d'abord les indices de similarité. L'indice de similarité est un indice qui prend des valeurs positives ; plus les deux objets se ressemblent, plus l'indice de similarité est grand, la similarité étant maximale lorsqu'une série est comparée à elle-même. Nous proposons ci-dessous une définition plus formelle.

Définition 1 : (Indice de similarité)
Un indice de similarité est une application qui vérifie les trois propriétés suivantes :

1. *positivité : s est une application $\Omega \times \Omega \to \mathbb{R}^+$*
2. *symétrie : s est symétrique : $\forall(\omega_1, \omega_2) \in \Omega \times \Omega,\ s(\omega_1, \omega_2) = s(\omega_2, \omega_1)$*
3. *$\forall(\omega_1, \omega_2) \in \Omega \times \Omega$ avec $\omega_1 \neq \omega_2,\ s(\omega_1, \omega_1) = s(\omega_2, \omega_2) > s(\omega_1, \omega_2)$*

Dans certains cas, il est préférable que la mesure de proximité prenne des valeurs minimales lorsque les séries sont proches. Nous introduisons dans ce but une nouvelle mesure de proximité d, opposée à l'indice de similarité.

Définition 2 : (Indice de dissimilarité)
Un indice de dissimilarité est une application qui vérifie les propriétés suivantes :

1. *positivité : d est une application $\Omega \times \Omega \to \mathbb{R}^+$*
2. *symétrie : d est symétrique : $\forall(\omega_1, \omega_2) \in \Omega \times \Omega,\ d(\omega_1, \omega_2) = d(\omega_2, \omega_1)$*
3. *identité : $\forall\omega_1 \in \Omega,\ d(\omega_1, \omega_1) = 0$*

Les indices de similarité et de dissimilarité sont deux notions semblables. En particulier, à tout indice de similarité s est associé l'indice de dissimilarité $d(\omega_1, \omega_2) = s(\omega_1, \omega_1) - s(\omega_1, \omega_2)$. Cependant, ces notions de similarité et de dissimilarité présentent certaines limites ; en particulier, deux objets différents peuvent avoir une dissimilarité nulle. Il faut, pour pallier cette limite, modifier l'axiome d'identité. En ajoutant à l'axiome ci-dessus sa réciproque, à savoir que la dissimilarité est nulle seulement quand un couple est comparé à lui-même, nous avons ainsi un axiome plus fort de séparation : on parle parfois de semi-métrique.
Pour quantifier l'écart entre deux objets, il est important que d'une part, plus deux objets se différencient, plus leur dissimilarité augmente, et d'autre part, que "le plus court chemin entre deux objets soit le chemin direct". Ces deux propriétés sont formalisées par l'inégalité triangulaire introduite ci-dessous. La notion de distance est ainsi définie de la manière

suivante.

Définition 3 : (Distance)
On appelle distance D sur un ensemble Ω une application $D : \Omega \times \Omega \to \mathbb{R}^+$ vérifiant les propriétés suivantes :

1. *symétrie :* $\forall \omega_1, \omega_2 \in \Omega \times \Omega, D(\omega_1, \omega_2) = D(\omega_2, \omega_1)$
2. *séparation :* $\forall \omega_1, \omega_2 \in \Omega \times \Omega, D(\omega_1, \omega_2) = 0 \Leftrightarrow \omega_1 = \omega_2$
3. *inégalité triangulaire :* $\forall \omega_1, \omega_2, \omega_3 \in \Omega, D(\omega_1, \omega_3) \leq D(\omega_1, \omega_2) + D(\omega_2, \omega_3)$

L'axiome de positivité est inclus dans les trois autres, . En effet,

$$\forall \omega_1, \omega_2, \underbrace{0 = D(\omega_1, \omega_1)}_{\text{Séparation}} \underbrace{\leq D(\omega_1, \omega_2) + D(\omega_2, \omega_1)}_{\text{Inégalité triangulaire}} \underbrace{= 2 * D(\omega_1, \omega_2)}_{\text{Symétrie}}$$

Une distance est donc un indice de dissimilarité vérifiant l'axiome de séparation et l'inégalité triangulaire.

Nous allons à présent donner des exemples de mesures de proximité dans le cadre des séquences temporelles. Les séquences sont des suites d'observations. Pour comparer les séquences, il faut comparer les observations deux à deux. Pour décider des observations à mettre en correspondance, nous introduisons la notion d'alignement, sur laquelle sont fondées la plupart des mesures de proximité usuelles dans le cadre des séquences. Nous formalisons à présent cette notion.

1.1.a Alignement de séquences

Une séquence est un vecteur d'observations qualitatives. Soit Σ un alphabet, i.e., un ensemble de symboles. On note $S^1 = (S_1^1, S_2^1, \ldots, S_T^1)$ une séquence de symboles de longueur T. Un alignement entre deux séquences est une manière de coupler les éléments des deux séquences en respectant certaines contraintes.

Définition 4 : (Alignement)

Un alignement de taille r est une application Φ de $\{1..r\}$ dans $\{1..T_1\} \times \{1..T_2\}$ vérifiant les propriétés suivantes :
- $\Phi(0) = (1; 1)$
- $\Phi(r) = (T_1; T_2)$
- $\forall t \in \{0 \ldots r-1\}$, $\Phi(t+1) = \Phi(t) + \Psi(t)$ *où* $\Psi(t) \in \mathcal{C}$ (*habituellement* $\mathcal{C} = \{(1; 0), (0; 1), (1; 1)\}$)

On notera en général $(t_1, t_2) = \Phi(t)$. Un alignement peut ainsi être vu comme la définition de deux vecteurs, un vecteur d'observations de S^1 et un vecteur d'observations de S^2.

$\mathcal{C}$ décrit un ensemble de contraintes pour l'alignement. On notera $\mathcal{A}_{\{T_1,T_2\}}$ l'ensemble des alignements entre deux séries de longueur T_1 et T_2. Un alignement est donc une façon de parcourir les instants de deux séries S^1 et S^2 en respectant certaines contraintes de monotonie (croissance des indices des observations), et de connexité (sauts d'au plus une unité entre les indices des observations).

A partir de la notion d'alignement décrite ci-dessus, nous pouvons définir quelques mesures de proximité entre séquences.

1.2 Distance de Levenshtein

Les travaux de Levenshtein (1965) ont contribué à définir une distance fondée sur les alignements. On appelle distance de Levenshtein le nombre minimal d'insertions, suppressions et substitutions pour passer d'une séquence à une autre. On qualifiera ces trois transformations de la séquence d'opérateurs d'édition.

Définition 5 : (Distance de Levenshtein (Levenshtein, 1965))
Soit S^1 et S^2 deux séquences de taille T_1 et T_2 sur un alphabet Σ. La distance de Levenshtein entre S^1 et S^2 correspond au nombre minimal d'opérateurs d'édition nécessaires pour transformer la séquence S^1 en S^2.

$$D(S^1, S^2) = \min_{\Phi \in \mathcal{A}} \Big(\#\{i \backslash S^1[i_1] = \text{" - "} \text{ ou } S^2[i_2] = \text{" - "}\} \Big) + \min_{\Phi \in \mathcal{A}} \Big(\#\{i \backslash S^1[i_1] \neq S^2[i_2]\} \Big)$$

où $\mathcal{A}$ représente l'ensemble des alignements entre les observations de S^1 et celles de S^2

Cette distance est très utilisée dans des applications réelles, en particulier pour l'analyse des séquences de nucléotides ; celles-ci sont sujettes à plusieurs types de mutations : les substitutions (échange entre deux nucléotides), les délétions (suppression d'un nucléotide), et les insertions (ajout d'un nucléotide). On y rajoute souvent les inversions de sous-séquences de nucléotides qui apparaissent quand un chromosome subit deux fractures et que le segment intermédiaire d'ADN est réinséré en sens inverse du sens initial, et également les translocations, qui correspondent à une insertion d'une sous-séquence dans une séquence de chromosome, parallèlement à la délétion de la même sous-séquence dans une séquence d'un autre chromosome. Pour établir la phylogénie entre les espèces vivantes, il est important de quantifier les modifications qu'ont subies les chaînes nucléotidiques.

1.3 Opérateurs d'édition

Dans les applications réelles, certains opérateurs d'édition sont plus fréquents que d'autres ; nous introduisons ici une fonction de score associée :

Définition 6 : (Fonction de score)
Pour formaliser insertions et suppressions, un caractère "blanc" est ajouté à l'alphabet. On note − ce nouveau caractère. Soit Σ un alphabet. On a, pour chaque opérateur, et pour chaque élément de l'alphabet augmenté $\Sigma \cup \{-\}$, un coût associé :

$$cout : \Sigma \cup \{-\} \times \Sigma \cup \{-\} \leftarrow \mathbb{R}$$

Par exemple, $cout(a, -)$ est le coût associé à la suppression du caractère a.
Nous définissons alors le score d'un alignement,

$$score(\Phi) = \sum_{i=1}^{r} \begin{cases} cout(a_{i_1}, b_{i_2}) \ si \ \Psi(t) = (1,1) \\ cout(a_{i_1}, -) \ si \ \Psi(t) = (1,0) \\ cout(-, b_{i_2}) \ si \ \Psi(t) = (0,1) \end{cases}$$

La fonction de score est une fonction qui associe à un alignement la somme des coûts des opérations d'édition. A partir de cette fonction de score, nous pouvons définir la distance d'édition généralisée qui recherche l'alignement minimisant le score.

Définition 7 : (Distance d'édition généralisée)
A chaque alignement, nous calculons le score associé à partir d'une fonction de coût donnée. La distance d'édition généralisée correspond au score minimal trouvé parmi tous les alignements.

Nous retrouvons, dans le cas où la fonction de coût est constante et égale à 1 pour tous les opérateurs d'édition, la distance de Levenshtein.

Notons que la qualification de cette fonction comme distance est un abus de langage. Selon la fonction de coût choisie, il ne s'agit pas d'une vraie distance. Nous développons dans la remarque suivante le contexte dans lequel la distance d'édition est une véritable distance.

Remarque 8 :

1. *La distance d'édition vérifie l'axiome de séparation si et seulement si pour tout $(a, b) \in \Sigma \cup \{-\}^2$, $cout(a, b) = 0 \Leftrightarrow a = b$*
2. *La distance d'édition est symétrique si et seulement si pour tout $(a, b) \in \Sigma \cup \{-\}^2$, $cout(a, b) = cout(b, a)$*
3. *La distance d'édition vérifie l'inégalité triangulaire si et seulement si pour tout $(a, b, c) \in \Sigma \cup \{-\}^3$, $cout(a, b) \leq cout(a, c) + cout(b, c)$*

Dans certaines applications, par exemple dans le contexte des séquences d'acides aminés, le choix de la fonction de coût est fondamental, certaines mutations étant plus fréquentes que d'autres. Des matrices de coût ont été définies de manière empirique, en particulier Pam (Dayhoff et Orcutt, 1979) et Blosum (Henikoff et Henikoff, 1992). Dans d'autres cas où la matrice de coût n'est pas connue, celle-ci peut faire l'objet d'un apprentissage préalable. Dans le cas de données structurées en arbre, (Boyer, 2011) propose un apprentissage probabiliste des coûts de la distance d'édition : par exemple, Hourai *et al.* (2004) propose une méthode

pour l'apprentissage de la fonction de coût fondée sur des méthodes bayésiennes.

Dans le cas où ne sont pas permises suppressions et insertions, nous définissons une distance d'édition particulière, appelée distance de Hamming. La distance de Hamming est une distance fréquente entre séquences. Elle se présente sous le même formalisme qu'une distance d'édition, ne considérant comme opérateurs d'édition que les substitutions. Les observations mises en correspondance sont celles apparaissant aux mêmes rangs.

1.3.a Distance de Hamming

Définition 9 : (Distance de Hamming (Hamming, 1950))
Soit S^1 et S^2 deux séquences de même longueur T sur un alphabet Σ. La distance de Hamming entre S^1 et S^2 correspond au nombre de différences entre les deux séquences.

$$D(S^1, S^2) = \#\{i \backslash S^1[i] \neq S^2[i]\} \tag{1}$$

L'opérateur D est une distance sur l'espace des séries temporelles. Les propriétés de symétrie et de séparation sont évidentes. L'inégalité triangulaire découle du fait que :

$$\begin{cases} S^1[i] = S^2[i] \\ S^2[i] = S^3[i] \end{cases} \Rightarrow S^1[i] = S^3[i]$$

dont la contraposée induit directement le résultat.

Remarque 10 :
La distance de Hamming est un cas particulier de la distance d'édition, avec un coût infini pour les suppressions et les insertions, et un coût de 1 pour chaque couple $(a, b) \in \Sigma^2$ $a \neq b$.

1.4 Plus longue sous-séquence commune

Deux séquences très différentes peuvent parfois être semblables sur une sous-partie de la série. Smith *et al.* (1981) introduisent l'algorithme de la plus longue sous-séquence commune (LCSS), un algorithme de programmation dynamique qui consiste à rechercher la plus longue sous-séquence commune en découpant les séquences en plusieurs préfixes (sous-séquences de petite taille), et allonger les préfixes tant qu'existe une similarité. A la fin, les différents préfixes sont recollés et forment la plus longue sous-séquence commune.

La recherche d'un alignement entre portions de la séquence est très important pour détecter localement des proximités. Par exemple, en biologie, deux séquences d'ADN partageant une sous-séquence commune peuvent être liées par un phénomène de "crossing-over", où les séquences nucléotidiques d'une partie d'un chromosome ont été insérées au cœur d'un autre chromosome. Il y a alors localement une forte proximité entre les deux séquences, qui n'apparaît pas forcément dans l'étude de l'ensemble des séquences à partir de la distance d'édition. Dans le cas de chromosomes ou de génomes circulaires, il est possible de généraliser toutes les proximités ou distances existant dans le cas linéaire (Demongeot *et al.*, 2009)

Nous avons présenté dans cette section un ensemble de mesures de proximité dans le cadre de séquences temporelles. Dans le cas où les observations ne sont plus symboliques, mais numériques, on définit des mesures de proximité de manière similaire. Nous nous intéressons donc dans la suite aux séries temporelles multivariées et aux mesures de proximité fondées sur la notion d'alignement.

2 Comparaison de séries temporelles multivariées

On appelle série temporelle une suite d'observations numériques évoluant dans le temps. Lorsque les observations sont vectorielles, on parle de séries temporelles multivariées.

On caractérise une série temporelle par une matrice de dimension $T \times p$, où T est le nombre d'instants qui définit la série et p le nombre de variables.

$$X = \begin{array}{c} \\ 1 \\ \vdots \\ T \end{array} \begin{array}{c} \begin{array}{ccc} X_1 & \dots & X_p \end{array} \\ \begin{pmatrix} S^1_{11} & \dots & S^1_{1p} \\ \vdots & S^1_{ij} & \vdots \\ S^1_{T1} & \dots & S^1_{Tp} \end{pmatrix} \end{array}$$

Dans le cas de plusieurs séries S^1,... S^n, nous considèrons que celles-ci s'expriment selon chaque variable et ont même longueur T. Nous exprimons les séries temporelles multivariées par une matrice à p colonnes et à $n \times T$ lignes, où sont regroupés les instants des différentes séries.

$$X = \begin{array}{c} \\ S^1 \\ \\ \\ \vdots \\ \vdots \\ \vdots \\ \\ S^n \\ \\ \end{array} \begin{array}{c} \begin{array}{ccc} X_1 & \dots & X_p \end{array} \\ \left(\begin{array}{ccc} S^1_{11} & \dots & S^1_{1p} \\ \vdots & S^1_{ij} & \vdots \\ S^1_{T1} & \dots & S^1_{Tp} \\ \hline & \vdots & \\ & \vdots & \\ & \vdots & \\ \hline S^n_{11} & \dots & S^n_{1p} \\ \vdots & S^1_{ij} & \vdots \\ S^n_{T1} & \dots & S^n_{Tp} \end{array}\right) \end{array}$$

2.1 Alignements de séries temporelles

Rappelons la définition 4 introduite précédemment. Un alignement de taille r est une application Φ de $\{1..r\}$ dans $\{1..T_1\} \times \{1..T_2\}$ vérifiant les propriétés suivantes :

- $\Phi(1) = (1;1)$
- $\Phi(r) = (T_1;T_2)$
- $\forall t \in \{1 \dots T\} \Phi(t+1) = \Phi(t) + \Psi(t)$ où $\Psi(t) \in \{(1;0),(0;1),(1;1)\}$

Notons que plusieurs hypothèses sont associées à cette définition des alignements :

- $\forall t_1 \in \{1 \dots T_1\}\ \exists r_1, t_2 / \Phi(r_1) = (t_1, t_2)$: les alignements ne présentent pas de sauts.
- $\forall r_1, r'_1, r_1 \leq r'_1 \Leftrightarrow t_1 \leq t'_1$ et $t_2 \leq t'_2$: la chronologie des observations est respectée.

Cette dernière hypothèse de monotonie apparaît en effet de manière fondamentale dans la définition de l'alignement. La fonction Ψ étant à valeurs dans $\{0;1\}$, le couple $\Phi(t+1)$ a ses deux coordonnées plus grandes que celles de $\Phi(t)$. L'alignement n'a pas de sauts, dans la mesure où la fonction Ψ est à valeurs dans 0 et 1, forçant l'alignement à être connexe (d'un seul tenant).

Remarque 11 : (Quelques propriétés des alignements)

- *La longueur des alignements est bornée :* $\min(T_1, T_2) \leq r \leq T_1 + T_2$
- *Un alignement de taille r contient :*
  - $T_1 + T_2 - r$ *déplacements selon les deux axes temporels* ($\Psi(t) = (1;1)$)
  - $r - T_2$ *déplacements selon l'axe de la première série* ($\Psi(t) = (1;0)$)
  - $r - T_1$ *déplacements selon l'axe de la seconde série* ($\Psi(t) = (0;1)$)
- *Dénombrement du nombre d'alignements possibles :*

$$\textit{Nombre d'alignements :} \sum_{r=\min(T_1,T_2)}^{T_1+T_2} C_r^{r-T_2} C_{T_2}^{r-T_1}$$

- *Les alignements sont connexes.*

La remarque précédente explicite un nombre important d'alignements possibles. En général, l'objectif est de choisir un alignement répondant de manière optimale à certaines contraintes ; on définit une fonction qui, à chaque alignement, associe un coût. Le problème d'optimisation associé aux alignements correspond à un problème de minimisation de cette fonction de coût.

Quand il n'y a pas d'ambiguïté sur la taille de la série, nous noterons l'ensemble des alignements possibles entre deux séries $\mathcal{A}$.
Attardons-nous sur un alignement qui prend beaucoup d'importance : l'alignement euclidien, noté Φ_{Euc}, qui est la fonction suivante :

Notation 1 : (Alignement euclidien)

$$\begin{array}{lll} \Phi_{Euc} : & [1,..T] & \to [1,..T] \times [1,..T] \\ & t & \mapsto (t,t) \end{array}$$

Dans la suite, nous proposons un formalisme utilisant la notion d'alignement pour définir les distances usuelles. Dans le cadre de ces approches, nous nous limiterons au cadre des alignements de l'ensemble des observations.

2.2 Un formalisme unifié pour une famille de mesures de proximité

Les mesures de proximité entre séries temporelles les plus couramment utilisées reposent toutes sur un même schéma. Elles sont dérivées de deux distances ou dissimilarités usuelles, "la distance Euclidienne" et la "Dynamic Time Warping". La première est la version la plus communément utilisée des distances de Minkowski, une famille de mesure entre vecteurs. Elle donne de très bons résultats pour la classification de séries où n'apparaissent pas de délais au sein des classes. La seconde a fait ses preuves dans beaucoup de domaines d'application,

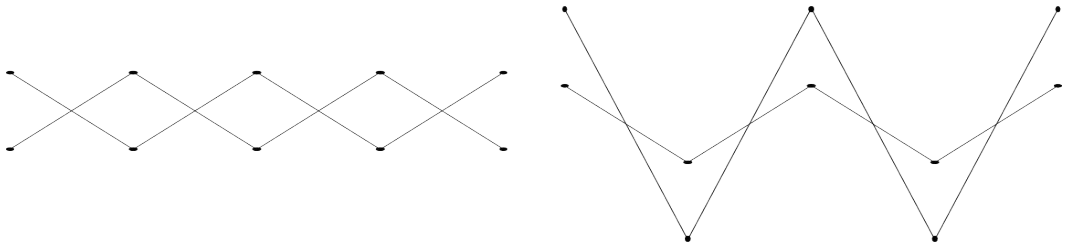


FIGURE 1 – Deux séries situées à la même distance

notamment en reconnaissance de signaux sonores et en biologie, et est réputée donner d'excellents résultats en classification, en particulier lorsque les séries se déduisent l'une de l'autre par des délais entre les instants des différentes séries. Elle est fondée sur la recherche d'alignements optimaux et s'inspire de l'idée introduite par Fréchet (1906) qui consiste à aligner les paires de séries de sorte à minimiser l'écart maximal entre elles. La distance de Fréchet est usuellement illustrée à partir de l'exemple suivant : un chien et son maître suivent tous les deux une trajectoire différente. Ils peuvent moduler leur vitesse ou faire des pauses, mais ne peuvent pas revenir en arrière. Quelle est la longueur minimale de la laisse ?

Prise en compte de la forme dans la définition d'une mesure de proximité Ces mesures de proximité sont fondées sur les écarts entre les valeurs prises par les séries et ne considèrent pas la dynamique au sein des séries. La figure 1 montre deux configurations : dans la partie gauche, deux séries s'opposent nettement, tandis que dans la partie droite, les deux séries vont dans le même sens, mais présentent des variations d'intensité.

Dans le contexte de séries temporelles, la configuration de gauche correspond à deux séries dont les comportements s'opposent fondamentalement, une décroissance de l'une correspondant à une croissance de l'autre ; la configuration de droite correspond à deux séries ayant subi un changement d'échelle. Cependant, les comportements globaux sont similaires : les périodes décroissantes et les périodes croissantes se correspondent.
La plupart du temps, deux séries temporelles partageant une même configuration sont considérées intuitivement comme proches, comme c'est le cas dans la configuration de droite, bien que les séries soient très différentes sur le plan des valeurs ; elles affichent des proximités sur le plan de la forme.

Une idée naturelle, lorsqu'on veut s'intéresser à la forme d'une série temporelle, est de passer dans l'espace des accroissements. Nous substituons aux séries S^l les séries de longueur $T-1$ et de terme général $V_i^l = S_{i+1}^l - S_i^l$. Les écarts sont calculés pour cette nouvelle série. Cependant, dans l'exemple précédent, au regard des accroissements pour les deux paires de séries de la figure 1, les couples d'instants se correspondant restent à une même distance. En effet, nous pouvons visualiser sur la figure 2 les accroissements des quatre séries.

Nous remarquons sur la figure 2 que, dans les deux configurations, les écarts entre les accroissements (traits pleins) sont en moyenne identiques. Dans ce contexte, le signe des accroissements est complètement ignoré.
Cependant, le signe joue un rôle fondamental : deux courbes croissantes ont un comportement plus similaire qu'une courbe croissante et une courbe décroissante, indépendamment des écarts entre leur pente.

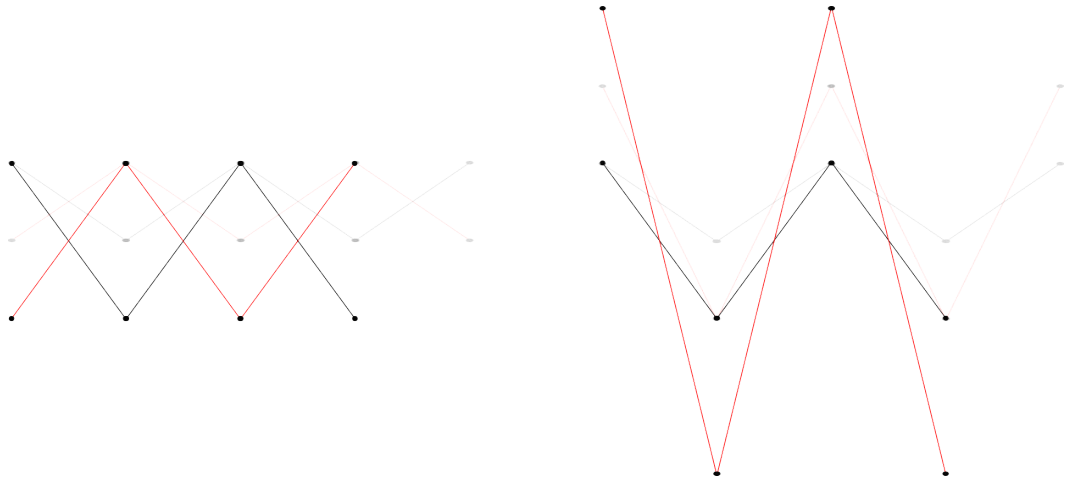


FIGURE 2 – Accroissements des séries précédentes

Nous allons donc nous intéresser dans la suite à quelques mesures de dissimilarité usuelles, certaines cherchant à tenir compte des différences de forme entre les séries.

Diallo (2010) et Douzal-Chouakria et Amblard (2011) proposent un formalisme unifiant toute une famille de mesures de proximité. Sont définis :

- un sous-ensemble R inclus dans l'ensemble des alignements $\mathcal{A}$.
- une fonction de coût $c(\Phi)$ d'un alignement Φ qui mesure l'écart en valeurs des observations couplées.
- une fonction de coût $co(\Phi)$ d'un alignement Φ qui mesure l'écart en forme des observations couplées.
- une fonction de couplage $f(c(\Phi), co(\Phi))$ de ces deux termes, qui rassemble les fonctions de coût en forme et en valeurs pour les observations des différents couples d'observations.

$$D(S^1, S^2) = \min_{\Phi \in R} f(c(\Phi), co(\Phi))$$

La métrique obtenue consiste à rechercher le minimum parmi tous les alignements d'une fonction de coût fondée à la fois sur les valeurs et sur la forme.

La fonction de coût la plus classique pour évaluer les écarts en valeur consiste à prendre la valeur absolue entre les deux observations. $c(\Phi)(S^1, S^2) = |S^1_\Phi(i)_1 - S^2_\Phi(i)_2|$

Pour la fonction de coût évaluant les écarts en forme, une première solution consiste à prendre la valeur absolue entre les accroissements.
Une seconde solution, proposée par Chouakria-Douzal (2003) consiste à utiliser la corrélation temporelle.

Corrélation temporelle A l'instar de la notion de dérivée pour les fonctions réelles, la notion d'évolution entre les différents instants est fondamentale pour une compréhension des séries. Certains travaux considèrent la forme des séries à partir des différences entre instants consécutifs. Les travaux menés sur les indices d'autocorrélation spatiale, initiés par Moran (1950) et Geary (1954), peuvent être étendus à une structure de contiguïté temporelle. Rappelons la formulation de la corrélation de Pearson à partir de l'ensemble des couples

d'observations.

$$COR(S^1, S^2) = \frac{\sum_{i,i'}(S_i^1 - S_{i'}^1)(S_i^2 - S_{i'}^2)}{\sqrt{\sum_{i,i'}(S_i^1 - S_{i'}^1)^2}\sqrt{\sum_{i,i'}(S_i^2 - S_{i'}^2)^2}} \quad (2)$$

Chouakria-Douzal (2003) introduit l'autocorrélation temporelle. Celle-ci limite le calcul de la corrélation aux instants voisins. La corrélation d'ordre r correspond à une corrélation limitée aux couples d'instants situés dans un voisinage proche. Elle se définit de la manière suivante :

Définition 12 : (Corrélation temporelle d'ordre r)

$$CORT(S^1, S^2) = \frac{\sum_{\substack{1\leq i\leq p-1 \\ |i-i'|<r}}(S_{i'}^1 - S_i^1)(S_{i'}^2 - S_i^2)}{\sqrt{\sum_{\substack{1\leq i\leq p-1 \\ |i-i'|<r}}(S_{i'}^1 - S_i^1)^2}\sqrt{\sum_{\substack{1\leq i\leq p-1 \\ |i-i'|<r}}(S_{i'}^2 - S_i^2)^2}} \quad (3)$$

Le cas $r = 1$ est classique, et calcule une corrélation entre observations successives.

$$CORT(S^1, S^2) = \frac{\sum_{i=1}^{p-1}(S_{i+1}^1 - S_i^1)(S_{i+1}^2 - S_i^2)}{\sqrt{\sum_{i=1}^{p-1}(S_{i+1}^1 - S_i^1)^2}\sqrt{\sum_{i=1}^{p-1}(S_{i+1}^2 - S_i^2)^2}} \quad (4)$$

La corrélation de Pearson et la corrélation temporelle CORT correspondent à des fonctions de coût en forme. En effet, l'écriture de la corrélation usuelle de Pearson à partir des couples d'observation et la formulation de la corrélation temporelle font toutes deux intervenir des couples de points pris dans les deux séries. La définition de la corrélation temporelle est en particulier compatible avec la notion d'alignement. Nous notons $lg(\Phi)$ la longueur de l'alignement Φ.

$$COR(S^1, S^2) = \frac{\sum_{i,i'}(S^1_{\Phi(i)_1} - S^1_{\Phi(i')_1})(S^2_{\Phi(i)_2} - S^2_{\Phi(i')_2})}{\sqrt{\sum_{i,i'}(S^1_{\Phi(i)_1} - S^1_{\Phi(i')_1})^2}\sqrt{\sum_{i,i'}(S^2_{\Phi(i)_2} - S^2_{\Phi(i')_2})^2}} \quad (5)$$

$$CORT(S^1, S^2) = \frac{\sum_{\substack{1\leq i\leq (lg(\Phi)-1) \\ |i-i'|<r}}(S^1_{\Phi(i')_1} - S^1_{\Phi(i)_1})(S^2_{\Phi(i')_2} - S^2_{\Phi(i)_2})}{\sqrt{\sum_{\substack{1\leq i\leq (lg(\Phi)-1) \\ |i-i'|<r}}(S^1_{\Phi(i')_1} - S^1_{\Phi(i)_1})^2}\sqrt{\sum_{\substack{1\leq i\leq (lg(\Phi)-1) \\ |i-i'|<r}}(S^2_{\Phi(i')_2} - S^2_{\Phi(i)_2})^2}} \quad (6)$$

2.2.a Distances usuelles

A partir de ces premières briques et de ce formalisme, nous pouvons construire toute une famille de distances. Nous retrouvons en particulier les distances usuelles.

Distances de Minkowski Les distances de Minkowski sont une suite de distances indexées par un réel k. La fonction de coût en valeurs est la valeur absolue de la différence $\delta(u, v) = |u-v|$, la fonction de coût en forme est constante égale à 1. Le sous-ensemble R des alignements

est réduit au seul alignement euclidien. La fonction de couplage est le produit cartésien $\left(\sum_{i=1}^{r} c(\Phi)(S^1,S^2)^k co(\Phi)(S^1,S^2)^k\right)^{\frac{1}{k}}$

La façon de coupler les différents instants au cœur de l'alignement est dictée par la valeur de k. Une valeur élevée de k donne plus d'importance aux écarts extrèmes. La distance $d^k(S^1,S^2)$ entre une série temporelle S^1 et une série S^2 est la distance suivante

$$d^k(S^1,S^2) = \left(\sum_{i=1}^{T}\left(S_i^1 - S_i^2\right)^k\right)^{\frac{1}{k}}$$

Les distances de Minkowski les plus classiques sont les distances d^1 (distance de Manhattan), d^2 (distance euclidienne), et $d^\infty = \max|S_i^l - S_i^{l'}|$ (distance de Chebychev). Une variante de ces distances est constituée par la distance $d^{p,q}$, définie à l'aide de la norme des espaces de Lorentz $L(p,q)$, dans laquelle l'élévation à la puissance p est d'abord utilisée, comme dans la distance d^p, suivie d'une élévation à la puissance $1/q$, comme dans la distance d^q (Coleman, 1982). Cette distance permet de privilégier, suivant les valeurs de p, les écarts extrêmes (grands ou petits), en compensant, par la valeur de $1/q$, des valeurs trop grandes ou trop petites de la somme.

Distances dans l'espace des phases Une dernière distance entre séries temporelles est la distance Δ_ϵ, dans l'espace $\Omega \times \Omega'$, dit espace des phases (ou des états), où $\Omega = \mathbb{R}^r$ est l'espace des séquences primaires $\{\omega_i\}$ et $\Omega' = \mathbb{R}^{r-1}$ l'espace des séquences d'accroissements $\{v_i = \omega_{i+1} - \omega_i\}$, dites aussi vitesses discrètes. Le point courant (ω, v) de $\Omega \times \Omega'$ est un ensemble de points $\{(\omega_i, v_i)\}_{i=1,\ldots,r-1}$ dans $\mathbb{R}^{2r-2}$, noté K. La distance Δ_ϵ entre deux points K et K' est définie par :

$$\Delta_\epsilon(K,K') = card(\{i\ ; d^2[(\omega_i,v_i),(\omega_i',v_i')] > \epsilon\})/(r-1)$$

c'est-à-dire par le pourcentage de «mismatches» (à ϵ près) dans l'espace des phases. La présence d'un décalage éventuel s (ou glissement indiciel) oblige à considérer la distance corrigée :

$$\Delta_\epsilon^C(K,K') = card(\{i\ ; d^2[(\omega_i,v_i),(\omega_{i+s}',v_{i+s}')] > \epsilon\})/(r-s-1)$$

Distances de Fréchet Les distances d'édition, définies dans le cadre de données symboliques, visent à chercher un chemin entre deux séquences, calculant la distance entre deux séries en autorisant suppressions, substitutions et insertions. Dans le contexte de variables continues, le pendant de la distance d'édition est la distance de Fréchet. C'est la dissimilarité correspondant à la distance infinie lorsqu'on considère tous les alignements possibles (Fréchet, 1906).

Définition 13 : (Distance de Fréchet)

Les fonctions de coût sont les mêmes que celles apparaissant dans le cadre des distances de Minkowski. L'ensemble des alignements possibles R est ici l'ensemble de tous les alignements $\mathcal{A}$. Le paramètre k est ici égal à ∞ ; la distance de Fréchet est le pendant de la distance de Chebychev lorsqu'on choisit l'alignement optimal.

$$d_{Frechet}(S^1, S^2) = \min_{\Phi \in \mathcal{A}(S^1,S^2)} (\max_{i=1}^{r} |S^1_{\Phi(i)1} - S^2_{\Phi(i)2}|) \quad (7)$$

La Dynamic Time Warping, une distance induite par des alignements La dynamic Time Warping (déformation dynamique temporelle) ou DTW (Sankoff et Kruskal, 1983; Sakoe et Chiba, 1978) recherche parmi tous les alignements possibles le chemin (non unique) qui minimise la distance de Manhattan des deux vecteurs alignés. Le développement des méthodes de programmation dynamique a été freiné à la fin des années 1980, avant de reprendre beaucoup de poids depuis quelques années, avec l'introduction de méthodes consistant à accélérer le processus.

Les fonctions de coût sont à nouveau les mêmes que celles apparaissant dans le cadre des distances de Minkowski, l'ensemble des alignements possibles R est ici l'ensemble de tous les alignements $\mathcal{A}$. Le paramètre k est égal à 1 pour la DTW. La DTW est donc le pendant de la distance de Manhattan.

$$\text{DTW}(S^1, S^2) = \min_{\Phi \in \mathcal{A}(S^1,S^2)} \sum_{i=1}^{r} |S^1_{\Phi(i)1} - S^2_{\Phi(i)2}| \quad (8)$$

Notation 2 : (Alignement DTW)
La DTW *est la distance associée à l'alignement minimisant l'écart entre les deux séries. On note* Φ_{DTW} *l'alignement associé.*

Notons que l'idée implicite est que les séries partagent un modèle commun et se déduisent l'une de l'autre par des petites variations en temps et en amplitude. En effet, l'alignement entre deux séries consiste à chercher la déformation du temps qui minimise l'écart entre les séries. Lorsque les séries ne partagent pas de structure proche, l'alignement a peu de sens.

Coût fondé sur les accroissements Pour obtenir des distances fondées sur la forme, certains travaux substituent à la série initiale la série des accroissements. C'est une idée naturelle lorsqu'on veut s'intéresser à la forme d'une série temporelle. Nous substituons aux séries S^l les séries de longueur $T-1$ et de terme général $V^l_i = S^l_{i+1} - S^l_i$. La distance est alors calculée sur la base des écarts en valeur entre les observations de cette nouvelle série.

$co(\Phi)(S^1, S^2) = c(\Phi)(V^1, V^2) = |S^2_{\Phi(i+1)2} - S^2_{\Phi(i)2} - S^l_{\Phi(i+1)1} + S^l_{\Phi(i)1}|$

Derivative DTW Cette méthode, proposée par Keogh et Pazzani (2001), consiste à appliquer l'algorithme de la DTW à la série des accroissements de terme général $(V^l_i)_{i \in 1..T-1} = (S^l_{i+1} - S^l_i)$. Cette méthode permet de prendre en compte la forme des séries de manière dynamique. (cf. section 2.2.a).

Corrélation temporelle La corrélation temporelle est une approche fondée sur les accroissements. Elle vaut 1 pour une situation où existe un lien fort entre deux séries, -1 pour un lien d'opposition entre deux séries, et 0 pour une situation neutre. L'indice de corrélation temporelle répond au problème de la forme des séries temporelles. Il permet de mettre en

avant des séries aux comportements similaires : il privilégie la tendance (croissance ou décroissance) à la valeur effective de la pente. De plus, du fait de la normalisation, il permet d'obtenir une valeur bornée. Les valeurs presque constantes ont un impact très faible sur le coefficient CORT.

Toutes les distances présentées ont leur spécificité et sont adaptées à un certain type de problème. Nous observons sur un exemple de deux séries temporelles les effets des différentes approches.

2.2.b Retour sur l'exemple précédent

Revenons sur les figures 1 et 2. La première montre deux configurations : dans la partie gauche, deux séries s'opposent nettement, tandis que dans la partie droite, les deux séries vont dans le même sens, mais présentent des variations d'intensité.

Dans le contexte de séries temporelles, la configuration de gauche correspond à deux séries dont les comportements s'opposent fondamentalement, une décroissance de l'une correspond à une croissance de l'autre ; la configuration de droite fait état d'un changement d'échelle. Cependant, les comportements globaux sont similaires. Les périodes décroissantes et les périodes croissantes se correspondent. Dans certains cadres, deux séries temporelles sont considérées proches si elles partagent une même configuration ; elles affichent des proximités en forme et non pas en valeurs.

Notons que la distance euclidienne appliquée aux accroissements considère les deux paires de séries de la figure 1 à nouveau à des distances égales. En effet, nous pouvons visualiser sur la figure 2 les accroissements des quatre séries.

Nous remarquons sur la figure 2 que dans les deux configurations, les écarts euclidiens entre les accroissements (traits pleins) sont identiques. Dans ce contexte, le signe des accroissements est complètement ignoré.
Cependant, le signe joue un rôle fondamental : deux courbes croissantes ont un comportement plus similaire qu'une courbe croissante et une courbe décroissante, indépendamment des écarts entre leur pente.

La corrélation temporelle permet de distinguer les deux séries. Elle vaut -1 pour la configuration de gauche, correspondant à une situation où existe un lien fort mais de sens opposé, et 1 pour la configuration de droite, correspondant à une situation où existe un lien fort et où les sens se correspondent. L'indice de corrélation temporelle répond au problème de la forme des séries temporelles. Il permet de distinguer les deux configurations qui sont intuitivement très différentes ; il privilégie la tendance (croissance ou décroissance) à la valeur effective de la pente.

La DTW recherche l'alignement qui minimise la distance. Dans l'exemple, les séries présentent un décalage d'une unité de temps. L'alignement qui minimise la distance est celui qui décale l'instant initial et l'instant final, et qui couple en décalé toutes les observations

$$\forall t \in \{2 \dots n-1\}\ \phi(t) = (t, t+1)$$

En fonction des applications, les méthodes fondées sur les alignements des instants induisent une proximité artificielle entre des séries sémantiquement très différentes, à l'instar de l'exemple précédent.

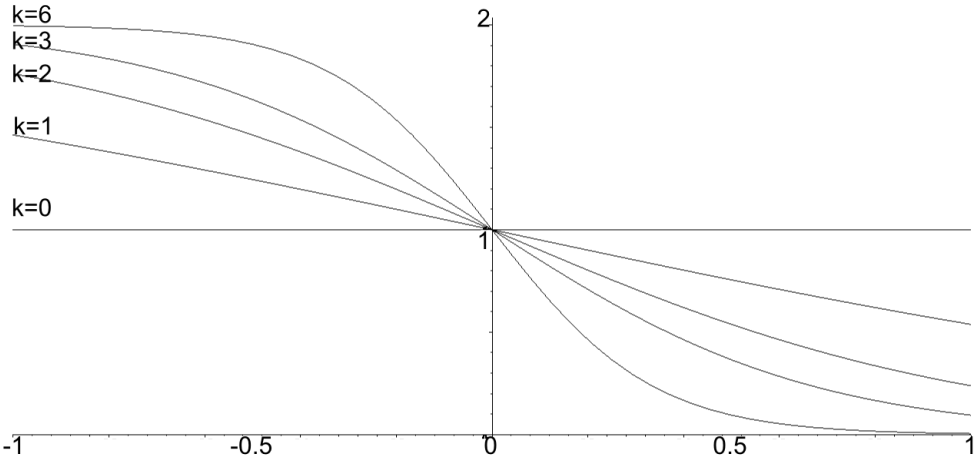


FIGURE 3 – Allure de la fonction f en fonction du paramètre k

2.2.c Distances proposant un compromis entre formes et valeurs

Les deux types de mesures de dissimilarités, fondées tantôt sur la forme, tantôt sur la valeur ne sont pas satisfaisantes vis-à-vis de la complémentarité des deux aspects. En fonction des applications, le coefficient de corrélation est parfois le plus adapté, parfois, c'est une distance fondée sur les valeurs, parfois encore, la mesure la plus adaptée est un compromis entre les deux mesures. Chouakria-Douzal et Nagabhushan (2007) proposent une métrique adaptative qui combine les deux approches à partir d'un paramètre amplifiant l'importance du facteur forme vis-à-vis du facteur valeur.

La mesure de dissimilarité choisie est fonction d'un paramètre k qui nuance le poids accordé à la forme.

Définition 14 : (Distance adaptative mêlant formes et valeurs)

$$d(S^1, S^2) = \frac{2\delta(S^1, S^2)}{1 + e^{k \times cort(S^1, S^2)}} = \delta(S^1, S^2) f(cort(S^1, S^2)) \qquad (9)$$

Cette métrique est adaptative. La valeur k fixe l'importance accordée au facteur forme. L'étude de la fonction $\frac{2}{1+exp^{kx}}$ sur $[-1, 1]$ permet d'observer l'incidence du paramètre k. La figure 3 montre l'allure de la fonction pour différentes valeurs de k. Pour $k = 0$, la fonction est constante égale à 1. L'impact de la forme est nul et tout le poids est porté par les valeurs. Pour $k \geq 6$, une forte corrélation temporelle donne une valeur très faible pour la fonction f, et donc une distance très faible, indépendamment des valeurs, tandis qu'une corrélation négative induit un poids fort.

La valeur du paramètre k est apprise, en général, pour chaque jeu de données, de sorte à minimiser la distance sur un échantillon d'apprentissage. La valeur trouvée est alors utilisée pour calculer les distances entre toutes paires de séries.

δ est une mesure de dissimilarité en valeur, par exemple la distance euclidienne. $CORT$

est définie à la section 2.2, et est une mesure de la dissimilarité en forme.

Alignement de l'ensemble des instants et notion d'appariement Les alignements présentés ci-dessus alignent l'ensemble des observations. Ils consistent à rechercher des liens au sein de tous les instants de la série. Dans certains cas, l'alignement d'un sous-ensemble d'observations a plus de sens, dans d'autres cas, il s'avère nécessaire de relâcher certaines des contraintes définissant les alignements. L'appariement de régions de forte similarité au sein de longues séquences est parfois préférable, lorsque les séries peuvent être très différentes dans leur ensemble. De même, des régions où apparaît un événement saillant ne se rangent pas toujours dans le même ordre au sein des séries. L'hypothèse de monotonie est parfois trop forte. Les approches sur des sous-ensembles sont souvent plus précises et préférables aux approches concernant l'ensemble des instants. L'identification des régions similaires est alors un défi supplémentaire à résoudre. Notons qu'une fois l'identification des régions d'intérêt faite, les méthodes d'alignements au sein des régions similaires peuvent s'effectuer de manière identique.

Parmi l'ensemble des mesures de proximité présentées ci-dessus, la distance euclidienne et la DTW sont les plus utilisées en général. Nous présentons dans la suite plusieurs variantes de la DTW portant principalement sur le choix de l'ensemble R des alignements et sur la pondération de la fonction de coût. Nous remarquerons que ces variantes sont générales et le formalisme présenté précédemment permet d'étendre facilement ces variantes à d'autres distances.

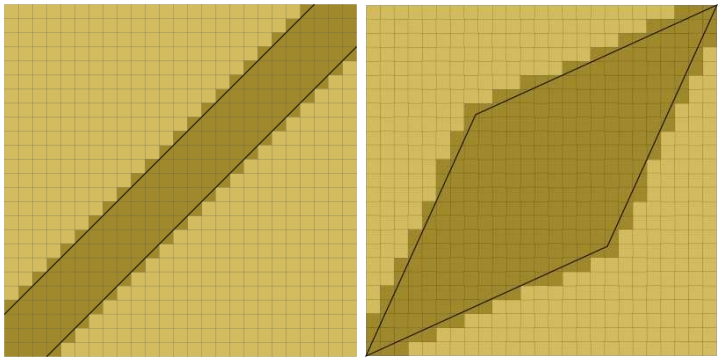
2.3 Variantes de la DTW

Les approches dynamiques telles que la DTW sont des méthodes robustes qui donnent de très bons résultats pour de nombreuses tâches de classification, et restent encore aujourd'hui référencées comme les méthodes les plus efficaces dans le cas général pour résoudre des problèmes de classification de type "k plus proches voisins" dans le cadre de séries temporelles (Yu et al. 2011). La DTW est utilisé comme point de comparaison de toute nouvelle méthode. Cependant, les alignements sélectionnés par la DTW peuvent coupler des zones totalement indépendantes. En effet, l'alignement n'a de sens que si les séries sont similaires. Autrement, la DWT consistant à rechercher un minimum, les alignements obtenus Φ_{DTW} peuvent n'avoir alors aucun sens.

Pour corriger les défauts de la méthode, des approches de régularisation des alignements ont été proposées. Nous présentons ici certaines de ces méthodes.

2.3.a Régularisation de la DTW

La régularisation de la DTW consiste à diminuer la complexité du modèle. Au lieu de chercher parmi tous les alignements celui qui optimise l'écart entre les séries, nous réduisons l'ensemble des alignements considérés. Restreindre le modèle en réduisant le nombre d'alignements possibles est a priori contre-intuitif, car la diminution de la complexité ne doit pas se faire au prix d'une diminution de la qualité des alignements. Cependant, dans de nombreuses applications, cette étape évite le sur-apprentissage qui peut conduire à des alignements très éloignés de l'objectif initial. Cette régularisation est toujours effectuée sur la



(a) Bande de Sakoe-Chiba (b) Parallélogramme d'Itakura

FIGURE 4 – Contraintes globales

base d'une connaissance a priori de la structure des données. Les techniques de régularisation les plus employées dans le cadre de la DTW ont été introduites par Sankoff et Kruskal (1983) et Sakoe et Chiba (1978) et se regroupent selon trois types de contraintes (Tomasi *et al.*, 2004) :

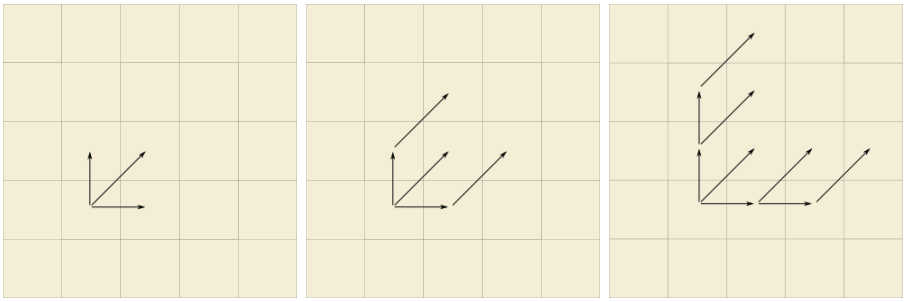
1. Contraindre la région dans laquelle peuvent apparaître les couples alignés.
2. Imposer des contraintes sur les pentes minimales et maximales des chemins.
3. Modifier l'impact de la longueur des chemins.

Fenêtre d'ajustement La première technique se traduit en général par la définition d'une fenêtre d'ajustement dans laquelle circulent les alignements appris. De telles fenêtres peuvent être de toutes sortes, les plus classiques sont les bandes de Sakoe-Chiba (Sakoe et Chiba, 1971) et le parallélogramme d'Itakura (Itakura, 1975) (cf. figure 4), qui correspondent à l'hypothèse intuitive selon laquelle un alignement entre deux séries ne doit pas s'éloigner trop significativement de la diagonale.

Ces deux matrices de contraintes peuvent être vues comme des matrices de pondération : lorsque l'arête sort de la fenêtre d'ajustement, le poids du couple est égal à $+\infty$. Ainsi, les contraintes globales peuvent se généraliser à un système de poids pénalisant tout écart à la diagonale.

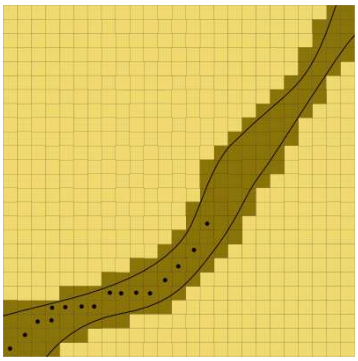
Modification des contraintes d'alignement Le second type de contraintes consiste à modifier la fonction Ψ apparaissant dans la définition d'un alignement (section 2.1). L'intérêt de ce type de contraintes au contraire des fenêtres d'ajustement est d'éviter que l'alignement ne se fixe, pour un nombre important d'instants, sur une observation donnée. (par exemple, $\Psi(t) = \Psi(t+1) = \ldots = \Psi(t+k) = (0,1)$, avec k grand)

Les trois configurations représentées à la figure 5 ont été proposées par Sakoe et Chiba (1978). La configuration (a) est le cas usuel proposé dans le cadre de la DTW, avec $\Psi(t) \in \{(1;1),(1;0),(0;1)\}$. Les configurations (b) ($\Psi(t) \in \{(1;1),(2;1),(1;2)\}$) et (c) ($\Psi(t) \in \{(1;1),(2;1),(3;1),(1;2),(1;3)\}$) forcent l'alignement à avancer à chaque itération d'un pas en avant. En particulier, un instant de la série S^1 ne peut être couplé dans ces deux configurations qu'à au plus un instant de la série S^2.



(a) (b) (c)

FIGURE 5 – Contraintes locales



(a) Bande du minimum local

FIGURE 6 – Contraintes de Rabiner *et al.*

Notons que ces contraintes induisent des bornes pour la pente des séries. Par exemple, la configuration (b) induit des pentes comprises entre $\frac{1}{2}$ et 2 dans la matrice de "Warping", la configuration (c) des pentes comprises entre $\frac{1}{3}$ et 3 et les contraintes d'alignement impliquent donc également des fenêtres d'ajustement.
Enfin, Rabiner *et al.* (1978) proposent de rechercher le chemin optimal à l'instant t+1 dans un voisinage de taille fixée pour le chemin déjà construit à l'instant t (cf. Figure 6).

2.3.b Pondération des couples et des chemins

Nous remarquons que la DTW présente un biais en faveur des chemins les plus courts, la dissimilarité se présentant comme une somme d'écarts où chaque terme représente un couple aligné. Ainsi, il est plus coûteux a priori de s'éloigner de la diagonale pour la contrainte locale usuelle (Figure 5 (a)) ; au contraire, dans le contexte des contraintes locales (b) et (c), les écarts à la diagonale sont favorisés. Dans certaines configurations, il faut d'éliminer ce biais. Pour cela, nous appliquons des pondérations ; celles-ci peuvent être de deux types : une première solution consiste à modifier le poids des chemins (par exemple division par la longueur du chemin pour obtenir l'écart moyen entre tous les couples alignés), une seconde solution consiste à affecter à chaque couple d'instants un poids intrinsèque. Nous pouvons remarquer que le biais lié à la longueur du chemin (cas de la figure (a)) est intéressant dans certaines applications, car il incite à favoriser les chemins les moins déformés.

Nous avons, dans ce qui précède, présenté plusieurs façons de calculer la proximité entre des séries temporelles. Nous avons observé que les mesures de proximité usuelles revenaient à la définition d'une mesure de proximité en forme, d'une mesure de proximité en valeur, et d'un ensemble d'alignements sur lesquels on cherche à optimiser une fonction coût calculée en fonction de ces deux mesures de proximité. La définition de l'ensemble des alignements conduit généralement à rapprocher les séries selon une approche paire à paire. A partir de ces alignements, et des mesures de proximité qui en découlent, telles que définies précédemment, nous abordons un problème de discrimination et de classification. Nous étudions comment les métriques définies précédemment peuvent être utilisées pour la classification de séries temporelles. Pour tenir compte de l'information au sein des classes, les travaux récents cherchent cependant à aligner les séries selon leurs caractéristiques communes à travers un alignement de l'ensemble des séries. Cette représentation est utile pour la discrimination de séries.

3 Classification de séries temporelles

Au sein d'un ensemble de séries, il est parfois important de découper l'ensemble en classes de séries, en fonction de similitudes au sein des classes, ou de différences entre les classes de l'ensemble. Dans certains cas, la composition des groupes est inconnue (classification dite non supervisée), et il faut proposer un découpage des classes. Dans d'autres cas, le découpage est connu (classification dite supervisée, par exemple des individus partageant une même modalité pour une variable explicative, telle une pathologie dans le cas de relevés physiologiques), et nous souhaitons expliquer ce qui discrimine les classes. De nombreuses méthodes de classification et de discrimination existent dans le cadre de données non temporelles. Dans le contexte des données temporelles, la tâche est plus complexe, car les méthodes doivent prendre en considération tant l'état que l'évolution des observations au cœur de la série. Dans le cadre de cette thèse, nous nous sommes intéressés, dans un contexte de discrimination, à la classification supervisée de séries temporelles. Warren Liao (2005) publie une synthèse des méthodes de classification appliquées aux séries temporelles. De nombreux travaux en classification reposent sur la définition d'une métrique. Dans le cadre de la classification supervisée, les classes peuvent alors être constituées selon un algorithme de type k-plus proches voisins, ou k-NN (de k Nearest Neighbor). Le principe de l'approche k-NN, introduite par Fix et Hodges (1951), consiste à affecter une nouvelle série dont l'étiquette est inconnue, à la classe à laquelle appartient la majorité des séries qui lui sont les plus proches au sens de la métrique, i.e., qui appartiennent aux k séries les plus proches de la nouvelle série. Cette méthode, de par sa simplicité et son efficacité, est encore très répandue. Toutes les métriques précédentes peuvent être utilisées au sein de ce processus. Notons que, dans le cadre de l'approche non supervisée, les méthodes usuelles fondées sur les algorithmes des k-moyennes ou PAM (de Partitioning Around Medoids, Kaufman *et al.* (1990)) nécessitent également l'utilisation de telles métriques. Il est indispensable de pouvoir calculer une dissimilarité entre les séries de l'échantillon et toute nouvelle série qui arrive. Le principe de l'algorithme consiste à calculer, pour chaque nouvelle série S, toutes les distances à chaque série de l'ensemble d'apprentissage $d(S, S^i),\ i \in \{1..k\}$, dont les étiquettes de classe sont connus. Les k plus proches voisins sont extraits et la classe majoritaire est désignée. En cas d'égalité, plusieurs solutions existent, consistant soit à affecter S à la classe correspondant à une valeur voisine de K (K-1 ou K+1...), ou de considérer une pondération des séries voisines en fonction de

leur distance, et de considérer la distance moyenne des séries les plus proches (Devroye *et al.*, 1996). Les performances de la méthode des k plus proches voisins peuvent être considérablement améliorées à travers des méthodes d'apprentissage. D'une part, la valeur optimale de k peut être apprise par validation croisée ; d'autre part, les métriques peuvent être apprises pour affiner la classification. Nous présentons dans la suite certaines méthodes d'apprentissage de métriques, en vue de la classification de séries temporelles par l'algorithme des k plus proches voisins.

3.1 Apprentissage de métriques en vue d'une classification *k*-NN

L'objectif d'une classification k-NN consiste à affecter chaque nouvel individu à la classe où se trouvent ses voisins les plus proches. Les métriques généralement définies entre séries temporelles ne distinguent pas les séries d'une même classe et les séries de classes différentes. La classification se fait sans a priori. L'apprentissage de métriques peut considérablement améliorer les résultats des classifications k-NN.

Distance de Mahalanobis (Mahalanobis, 1936) La définition de la distance euclidienne est liée à des écarts entre couples d'observations. Une transformation linéaire de l'espace peut être opérée au préalable pour favoriser les dimensions dans lesquelles les variables sont les plus discriminantes. La transformation linéaire se traduit par une matrice $\mathcal{L}$.

La matrice $A = {}^t\mathcal{L}\mathcal{L}$ est symétrique et semi-définie positive. On en déduit une métrique d_A introduite par Mahalanobis (1936)

Définition 15 : (Distance de Mahalanobis)
On définit la distance de Mahalanobis par $d_A(X,Y) = {}^t(X-Y)A(X-Y)$

Un cas particulier de cette distance consiste à prendre comme matrice A une matrice diagonale de terme général σ_i^2, la variance de l'observation X_i. D'autres travaux cherchent à apprendre la matrice A associée à une distance de Mahalanobis qui minimise les distances entre les séries d'une même classe et qui maximise les distances entre deux séries de classes différentes (Xing *et al.*, 2002).

Hastie et Tibshirani (1996) proposent d'apprendre une métrique discriminante : sur la base du quotient des variances inter B sur intra W, ils définissent la métrique $\Sigma \cong W^{-1}BW^{-1}$ qui construit localement un espace transformé pour le calcul des k plus proches voisins.

Largest margin NN Weinberger *et al.* (2006) ont introduit la notion d'imposteurs. On définit un imposteur comme un individu parmi les k plus proches voisins, n'appartenant pas à la classe de la série. Idéalement, l'objectif est de n'avoir aucun imposteur parmi les k plus proches voisins. Les travaux consistent à apprendre la matrice A associée à une distance de Mahalanobis qui éloigne les imposteurs et construit une marge entre les imposteurs et les voisins.

Des travaux ont été conduits sur des similarités et non plus des distances. En particulier, Qamar *et al.* (2008) généralisent la similarité du cosinus à une matrice semi-définie positive quelconque et proposent une méthode d'apprentissage d'un sous-espace dans lequel une classification k-plus proches voisins donne des résultats optimaux. L'espace optimal est recherché

par une méthode dérivée de la méthode du perceptron, puis la matrice A est projetée dans l'espace des matrices semi-définies positives par annulation des valeurs propres négatives.

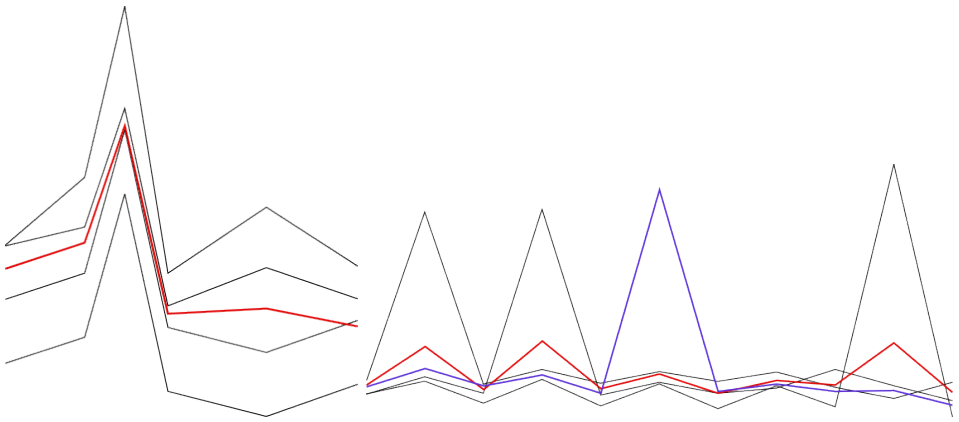
3.2 Classification non supervisée et définition d'un prototype

A l'image de l'algorithme des k-médoïdes, la classification non-supervisée s'apparente souvent à la recherche d'un individu central au sein de chaque classe. Dans le cadre de séries temporelles, un individu central correspond à un profil type, partagé par les séries de la classe. Dans le cadre d'espaces euclidiens usuels, la notion de barycentre est intrinsèque, et il est facile de définir un centroïde de la classe. Dans le cadre temporel, la définition d'un centre de classe est délicate. Nous considérons dans ce paragraphe plusieurs approches proposées pour définir un profil central.

Les méthodes les plus naïves consistent à considérer une série temporelle de longueur T comme un vecteur de $\mathbb{R}^T$ et d'appliquer les méthodes usuelles pour des données multidimensionnelles (moyenne instants par instants, médiane...). Ceci néglige le lien qui peut exister entre les instants (notamment la chronologie temporelle) et la forme des séries. Notons que ces méthodes ne s'appliquent que dans le cas de séries temporelles de même longueur. Warren Liao (2005) répartissent en trois catégories les méthodes de classification non supervisées de séries temporelles. Les **méthodes fondées sur les données brutes**, à l'instar de la méthode ci-dessus, consistent à manipuler les séries pour se ramener à un problème statique. Nous substituons pour ce faire à la dissimilarité de l'approche statique, une dissimilarité adaptée aux séries temporelles. Ces méthodes englobent notamment la méthode des k-moyennes, appliquée à des métriques temporelles ; par exemple Wilpon et Rabiner (1985) appliquent la méthode à la DTW. La seconde approche regroupe les **méthodes fondées sur l'extraction d'un nombre réduit de caractéristiques** des séries. Il englobe en particulier toutes les méthodes fondées sur la notion de transformée de Fourier discrète (Cooley et Tukey, 1965; Agrawal *et al.*, 1993), les corrélations croisées (Goutte *et al.*, 1999), la décomposition en ondelettes (Chan et Fu, 1999) et des Analyses en Composantes Principales (ACP) adaptées à des séries temporelles (Yang et Shahabi, 2004). Le troisième type de méthode consiste en des **approches fondées sur la notion de modèle**. On y retrouve les méthodes de types Modèles de Markov cachés. Ces dernières méthodes cherchent à trouver un modèle qui explique les séries.

Prototype de séries temporelles fondé sur les alignements DTW La recherche d'un prototype associé à un ensemble de séries alignées est un problème important dans le contexte de la classification non supervisée. L'absence de produit scalaire naturel pour des séries temporelles, qui empêche la définition d'un barycentre, rend délicate la tâche de définir une série type au sein d'une classe.

Comme nous pouvons le voir sur la figure 7 (a), lorsque les séries des classes sont proches et se correspondent à chaque instant, la série $(\mathbb{S}_i)_{i\in\{1..T\}}$ (en rouge) est un bon résumé de l'ensemble des séries. Cependant, quand les séries subissent des délais et sont alignées par la DTW, la notion de barycentre instant par instant n'a aucun sens, comme l'illustre la figure 7 (b)). Cette moyenne ne reflète le comportement d'aucune des séries. De nombreux travaux se sont penchés sur le problème de moyennage d'un ensemble de séries dans le cadre des alignements obtenus par la DTW.



(a) Séries aux comportements similaires

(b) Séries aux comportements différents

FIGURE 7 – Barycentre instant par instant des séries

Gaffney et Smyth (2005) proposent d'apprendre les prototypes sur la base d'un modèle probabiliste tenant compte des alignements. Le prototype prend en considération tant les écarts en valeurs que les délais. Srisai et Ratanamahatana (2009) proposent de calculer une moyenne des séries tenant compte des délais moyens. Les événements saillants sont alignés et la série moyenne comporte ces événements apparaissant à une date qui est la moyenne des dates des séries (cf. courbe bleue figure 7 (b)). Cependant, ce type d'approche crée une série temporelle en décalage avec toutes les séries de l'ensemble. Nous voyons que la série bleue comporte un événement saillant apparaissant à une date où rien ne se produit dans les séries initiales. Le prototype ainsi créé n'a aucun lien avec les séries qu'il représente.

D'autres approches consistent à rassembler les séries deux à deux par un alignement DTW. Cependant ces approches dépendent de l'ordre des séries. Pour calculer la moyenne des séries, Abdulla *et al.* (2003) recherchent d'abord un médoïde, i.e., une série qui minimise la DTW moyenne avec toutes les séries de la classe. Toutes les séries sont alignées sur ce médoïde par une DTW, puis la série moyenne est calculée sur la base des séries alignées sur ce médoïde. La série moyenne a la même taille que le médoïde.
Hautamaki *et al.* (2008) se fondent sur cette notion et réitèrent le processus, en réalignant les séries sur ce prototype, et en recalculant à chaque itération le prototype. Petitjean *et al.* (2011) proposent un algorithme itératif qui vise à modifier une série initiale pour minimiser les écarts quadratiques selon l'alignement (DTW) à toutes les séries dont nous souhaitons calculer la moyenne.

Problèmes inhérents à la notion d'alignement paire à paire Les travaux présentés ci-dessus reposent sur des métriques définies à partir de la notion d'alignement. Comme il est noté chez Listgarten (2007), il est important de remarquer que le problème de la recherche d'alignement, efficace pour la classification de séries temporelles, n'apporte pas une compréhension de la structure qui lie un ensemble de séries. La recherche d'alignements vise à faire coïncider deux éléments ; par exemple, dans le cas de la reconnaissance vocale, l'alignement permet de voir si les personnes expriment la même chose deux à deux, mais ne permet pas de dégager un consensus au sein des différentes séries, de définir une diction type. Un point important pour l'exploration et la discrimination de séries réside dans la recherche de structures

d'appariements unifiées entre toutes les séries. une série "moyenne", qui partage les éléments communs entre toutes les séries initiales. La recherche d'alignements tenant compte de l'ensemble des séries et non limité aux paires de séries est un problème difficile, très important dans un contexte de discrimination.

3.3 Alignements multiples

La notion d'alignements multiples est une notion qui s'est beaucoup développée avec les avancées des outils informatiques. Un alignement multiple est un tableau constitué d'insertions, suppressions et substitutions entre toutes les séquences, de sorte à optimiser un score global associé aux séries. L'alignement paire à paire présenté précédemment en est un cas particulier, optimisant la similarité entre deux séries ou deux séquences. L'objectif de cette section est de présenter les différentes manières d'appréhender les alignements entre paires de séries pour créer un alignement multiple. Dans le cadre des séquences, la recherche d'alignements multiples est un problème très important de ces trente dernières années ; il trouve en particulier des applications réelles, notamment en biologie pour les alignements de séquences de nucléotides ou de peptides. La plupart des solutions présentées ci-dessous ont été proposées dans ce cadre. L'alignement de multiples séries apporte davantage d'informations structurelles pour des tâches de classification, de détection de motifs, de prédiction de structure et de compréhension des relations entre les séries. Dans le cadre d'un article de synthèse, Notredame (2002) fait un état de l'art des méthodes d'alignements multiples, mettant en avant trois types d'approches pour attaquer ce problème, les approches dites exactes, les approches progressives et les approches itératives.

- Les approches exactes consistent à calculer un alignement optimal ; cependant, les problèmes d'alignement sont en général trop coûteux en ce qui concerne la complexité pour ce calcul, et les approches exactes sont limitées à un petit nombre de séries.
- Les approches progressives consistent à construire l'alignement multiple en ajoutant un par un les alignements paire à paire ; ces méthodes, bien qu'étant les plus utilisées, n'assurent aucunement l'optimalité de l'alignement obtenu.
- Les approches itératives consistent à raffiner l'alignement tout au long d'un processus itératif (de manière déterministe ou stochastique).

Dans la pratique, les méthodes reposent souvent sur une combinaison d'aspects progressifs et itératifs. La fonction objectif sur laquelle reposent les alignements multiples doit être définie correctement pour rechercher des "éléments communs" entre l'ensemble des séries. La définition de la fonction objectif n'est de loin pas triviale, et conduit à un ensemble de méthodes très variées pour l'apprentissage d'alignements multiples. Nous détaillons dans la suite les approches les plus courantes.

3.3.a Approches exactes

Les approches de programmation dynamique fondées sur les fonctions de coût peuvent se généraliser au cadre d'un ensemble de séries temporelles. L'objectif est de maximiser la similarité de l'ensemble des paires de séries. Pour cela, la méthode naïve consiste à considérer l'ensemble des séquences comme un tableau de grande dimension, où chaque séquence définit une dimension, puis à étendre l'algorithme de programmation dynamique classique de Needleman *et al.* (1970) à un espace multidimensionnel. Un alignement dans cet espace

de grande dimension est alors obtenu. Cependant, ces méthodes ne passent pas à l'échelle pour des questions calculatoires, la complexité augmentant exponentiellement en fonction du nombre de séquences. Elles sont en pratique impossibles à mettre en œuvre dès lors qu'on étudie plus d'une dizaine de séquences.

3.3.b Approches progressives

Ces méthodes sont les plus fréquentes. Elles construisent un alignement multiple en combinant les alignements paire à paire selon des distances entre séries (par exemple un arbre phylogénétique ou une matrice de distance). Ces algorithmes commencent par coupler les séries les plus proches, puis progressivement couplent les séries les plus éloignées (Hogeweg et Hesper, 1984; Feng et Doolittle, 1987). Les méthodes progressives sont souvent utilisées pour leur rapidité, pouvant aligner jusqu'à plusieurs milliers de séquences, mais ces approches présentent le défaut de ne pas revenir sur les alignements déjà appris (Notredame, 2002). Le fait de choisir un ordre pour aligner les séries n'assure pas la convergence vers des optima globaux. Le résultat dépend fortement de l'ordre des séquences. En particulier, quand les séquences présentent toutes des profils assez éloignés, la propagation des erreurs conduit à des alignements multiples inadaptés (Duret et Abdeddaim, 2000).

Les méthodes progressives actuelles ajoutent un second système de poids, lié à la proximité phylogénétique, pour corriger ce problème. Les différentes variantes proposées pour résoudre le problème de la recherche d'alignements multiples par méthodes progressives résident dans la construction du dendrogramme guidant l'approche paire à paire.

Citons par exemple COFFEE (Notredame *et al.*, 1998) et sa version améliorée T-COFFEE (Notredame *et al.*, 2000) basés sur la notion de cohérence.

3.3.c Approches itératives

Les approches itératives consistent à construire une solution de manière itérative, en améliorant à chaque étape une solution existante. Les méthodes itératives se classent en deux catégories, les algorithmes stochastiques et les algorithmes non stochastiques. Les premiers peuvent être fondés sur le principe du recuit simulé (opérer des modifications de manière aléatoire et conserver la modification en cas d'amélioration) ou sur des algorithmes génétiques (des alignements multiples générés aléatoirement subissent des "mutations" et des "croisements" ; les plus adaptés survivent, tandis que les autres périssent). Les deux approches sont parfois couplées, dans le cadre de la recherche d'alignements (Cai *et al.*, 2000).
Les méthodes non stochastiques visent à poursuivre à chaque itération l'alignement multiple, par un processus d'alignement classique paire à paire.

Citons quelques exemples de méthodes itératives : MUSTA (MUltiple STructure Alignment, Leibowitz *et al.* (2001)) recherche la plus grande sous-structure commune à toutes les séries ; MASS (Multiple Alignment by Secondary Structure, Dror *et al.* (2003)) et MultiProt (Shatsky *et al.*, 2002) construisent l'alignement multiple sans aligner nécessairement toutes les séries. Nous présentons dans la suite une méthode itérative d'alignement de séquences qui propose un alignement limité à certains instants.

Une méthode itérative : DIALIGN DIALIGN est une méthode d'alignements paire à paire qui s'étend à l'apprentissage d'alignements multiples. Chaque alignement se caractérise

par une matrice couplant les positions d'une séquence par rapport à l'autre. La méthode DIALIGN (Morgenstern *et al.*, 1996) recherche des alignements constitués de diagonales qui sont des blocs carrés inclus dans la matrice et correspondant à un alignement euclidien des deux sous-séquences. Elle considère au départ toutes les diagonales possibles entre les deux séries et leur affecte un score. Les diagonales sont alors sélectionnées sur la base d'un critère de cohérence, afin de maximiser ce score. Le critère de cohérence entre deux diagonales est un critère de monotonie.
Dans le cadre de l'alignement multiple, le critère de cohérence est étendu à l'ensemble des séries. Le mode de construction consiste à prendre au départ une structure d'alignement vidée de toutes les diagonales et à construire l'alignement multiple en incluant les diagonales par ordre de score, en s'assurant que chacune respecte le nouveau critère de cohérence. L'alignement est multiple en ce que toutes les diagonales sont incluses au regard de l'ensemble des diagonales déjà présentes.
Le parcours exhaustif de toutes les diagonales est très coûteux. Pour accélérer le processus , la méthode DIALIGN peut être couplée à une autre méthode d'alignements appelée CHAOS (Brudno *et al.*, 2004). CHAOS recherche les points d'ancrage des diagonales de DIALIGN sous la forme de graines, qui sont des diagonales de petite taille (la taille est un paramètre de la méthode). Les graines peuvent être reliées si elles sont suffisamment proches et respectent les contraintes de cohérence.

Pour terminer cette approche bibliographique, nous présentons quelques méthodes utilisées également pour la classification et la discrimination de séries temporelles.

3.4 Autres méthodes

De nombreuses méthodes utilisées classiquement pour la classification ont été étendues à la structure particulière des séries temporelles. C'est notamment le cas des modèles de Markov et des méthodes à noyaux. Nous présentons dans cette dernière partie quelques travaux allant dans ce sens.

3.4.a Modèles de Markov cachés (HMM)

Certaines méthodes visant à analyser les séries temporelles consistent à utiliser des modèles de Markov cachés ou HMM, de l'anglais Hidden Markov Model.

Principe Les HMM sont des méthodes statistiques, qui construisent des automates probabilistes formalisés par des chaînes de Markov. Ils sont définis par un ensemble d'états et des probabilités de transitions, et consistent à capturer la structure de la séquence de probabilité maximale associée à une série. L'état réel du processus est caché. Dans le cadre des séries temporelles, on considère un type de transition particulier. Owsley *et al.* (1997) introduisent

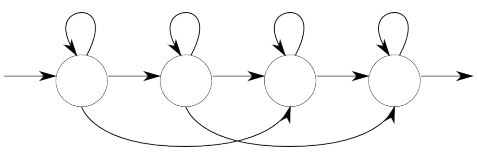


FIGURE 8 – Type de transitions dans le cadre des HMM

l'idée de définir des classes par un HMM, et utilisent les k-moyennes en vue d'une classification non supervisée. Li et Biswas (1999) proposent une méthode d'apprentissage des paramètres.

Limites Les méthodes HMM posent des problèmes particuliers dans le cadre de notre contexte. La propriété de Markov est très contraignante et impose des restrictions (indépendance des états passés). De même, les probabilités de transition reposent souvent sur l'hypothèse gaussienne. Enfin, la difficulté de l'apprentissage des états initiaux et des probabilités de transition peut rendre les modèles HMM moins compétitifs que la DTW (Ravinder, 2010). Les alignements de la Dynamic Time Warping peuvent parfois conduire à des erreurs. Les modèles de Markov cachés sont des méthodes plus performantes, mais nécessitent une information a priori. Plusieurs auteurs ont ainsi cherché à coupler les méthodes HMM avec la DTW (Oates *et al.*, 1999, 2001; Hu *et al.*, 2006)

Utilisation de l'information contenue dans toutes les classes La plupart des méthodes de classification consistent à étudier l'adéquation d'un individu à une classe. La notion d'imposteurs consiste à utiliser l'information contenue dans les classes voisines, pour construire une méthode de classification. Cependant, les problèmes d'alignement détaillés précédemment sont fondés uniquement sur la structure au sein de la classe.
Dans certains cas, une classe est caractérisée par le fait de se différencier avec un événement d'une autre classe. Listgarten (2007) propose dans sa thèse d'améliorer l'apprentissage des alignements, généralement construits par le rapprochement de séries proches, en impliquant les différences entre séries de classes différentes. En général, la détection des différences est traitée a posteriori, une fois les alignements connus.

3.4.b SVM : Méthodes fondées sur les noyaux

Les méthodes fondées sur les noyaux consistent à définir une fonction Kernel, qui consiste en une transformation non nécessairement linéaire des données pour se placer dans un espace où les classes sont séparées. Plusieurs travaux ont été menés pour étendre les travaux sur les noyaux au cadre des séries temporelles. Citons notamment les travaux de Haussler (1999) sur les noyaux appliquées à des structures de chaînes et ceux de Vert *et al.* (2004); Mahé et Vert (2009) sur les noyaux fondés sur des alignements locaux, qui utilisent l'idée de convolution et étudient l'ensemble des sous-segments de la chaîne ; les travaux de Cuturi *et al.* (2007) proposent un noyau fondé sur les opérations élémentaires de la DTW par un passage à l'exponentielle du score optimal obtenu par la DTW. Yang et Shahabi (2005) définissent un noyau fondé sur l'ACP du tableau des séries, avec une projection dans l'espace des composantes principales. Enfin, Rüping (2001) compare des approches fondées sur plusieurs noyaux appliqués aux séries temporelles (Linear kernel, Radial Basis, Fourier ...).

Conclusion

Nous avons introduit dans ce chapitre la notion de séquences temporelles et de séries temporelles et nous avons étudié les mesures de proximité usuelles qui en découlent. La plupart de ces mesures de proximité reposent sur un alignement des instants des deux séries. Pour la discrimination d'ensembles de séries temporelles, les alignements paire à paire sont

souvent inadaptés. Nous avons présenté la notion d'alignements multiples pour lier les instants d'un ensemble de séries. Enfin, nous avons exploré plusieurs méthodes pour la classification de celles-ci, fondées sur ces alignements. En vue d'une amélioration de la métrique, nous souhaitons apprendre les alignements multiples pour maximiser la discrimination. Pour cela, nous considérons le lien entre les différentes séries comme une structure de voisinage. Nous introduisons ainsi dans le prochain chapitre la variance associée à une structure de contiguïté en vue de la définition d'un critère d'optimisation pour l'apprentissage.

Chapitre 2

Analyse de l'interdépendance des données

Nous avons fait dans le chapitre précédent un état de l'art des mesures de comparaison entre paires de séries temporelles et nous nous sommes intéressés à la notion d'alignements multiples. En vue de la discrimination de séries temporelles, nous souhaitons apprendre des alignements optimisant un critère de discrimination. La variance est un indicateur souvent utilisé pour les problèmes de séparation et de différentiation. A partir d'une première étude des indices de Moran et de Geary, et des travaux sur la variance de données contiguës, nous étendons les travaux effectués dans le cadre de données contiguës à un ensemble puis à une partition de séries temporelles. Nous considérons le lien entre les différentes séries comme une structure particulière de voisinage.

En général, le tableau de données étudiées dans une analyse est supposé homogène. Il n'existe pas de dépendances a priori entre les individus et entre les variables. Or, il est fréquent que le tableau de données soit déjà structuré. C'est le cas par exemple des données géographiques et temporelles où une structure existe entre les observations. Il est, dans ce contexte, fréquent d'utiliser des méthodes spécifiquement adaptées à ce type de données. Ces méthodes d'analyse se regroupent sous le terme d'analyse de la contiguïté. Dans le cadre de notre travail sur les séries temporelles, nous nous intéressons donc au cadre général des données contiguës. Nous allons donc dans ce chapitre nous intéresser à l'extension de la variance définie dans le cadre de données contiguës à des ensemble et à des partitions de séries temporelles.

L'analyse de la contiguïté regroupe des extensions de méthodes d'analyses exploratoires classiques à des données contiguës. Ces approches proposées par Lebart (1969), par Wartenberg (1985) et Banet et Lebart (1984) ont pour point de départ les indices d'autocorrélation spatiale de Moran (1950) et de Geary (1954). Elles visent à généraliser les notions de variance et de covariance à des données liées par une structure de contiguïté non nécessairement temporelle. Mom (1988) introduit une notion de variance visant à généraliser l'analyse factorielle discriminante à une notion plus générale de classe d'objet fondée sur les liens de voisinage, à partir d'une métrique spécifique.

Introduction de la notion de contiguïté La contiguïté des observations se caractérise par un système de poids associé à chaque couple des séries observées. La figure 2 présente différents types de structures de contiguïté usuels. Les plus classiques sont les structures de contiguïté spatiales (deux régions sont voisines si elles partagent des frontières communes) et temporelles (deux instants sont voisins s'ils correspondent à des instants consécutifs). D'autres types de lien peuvent encore apparaître en fonction des observations, par exemple, deux régions sont proches si elles partagent des valeurs proches. Dans ce contexte, Lebart

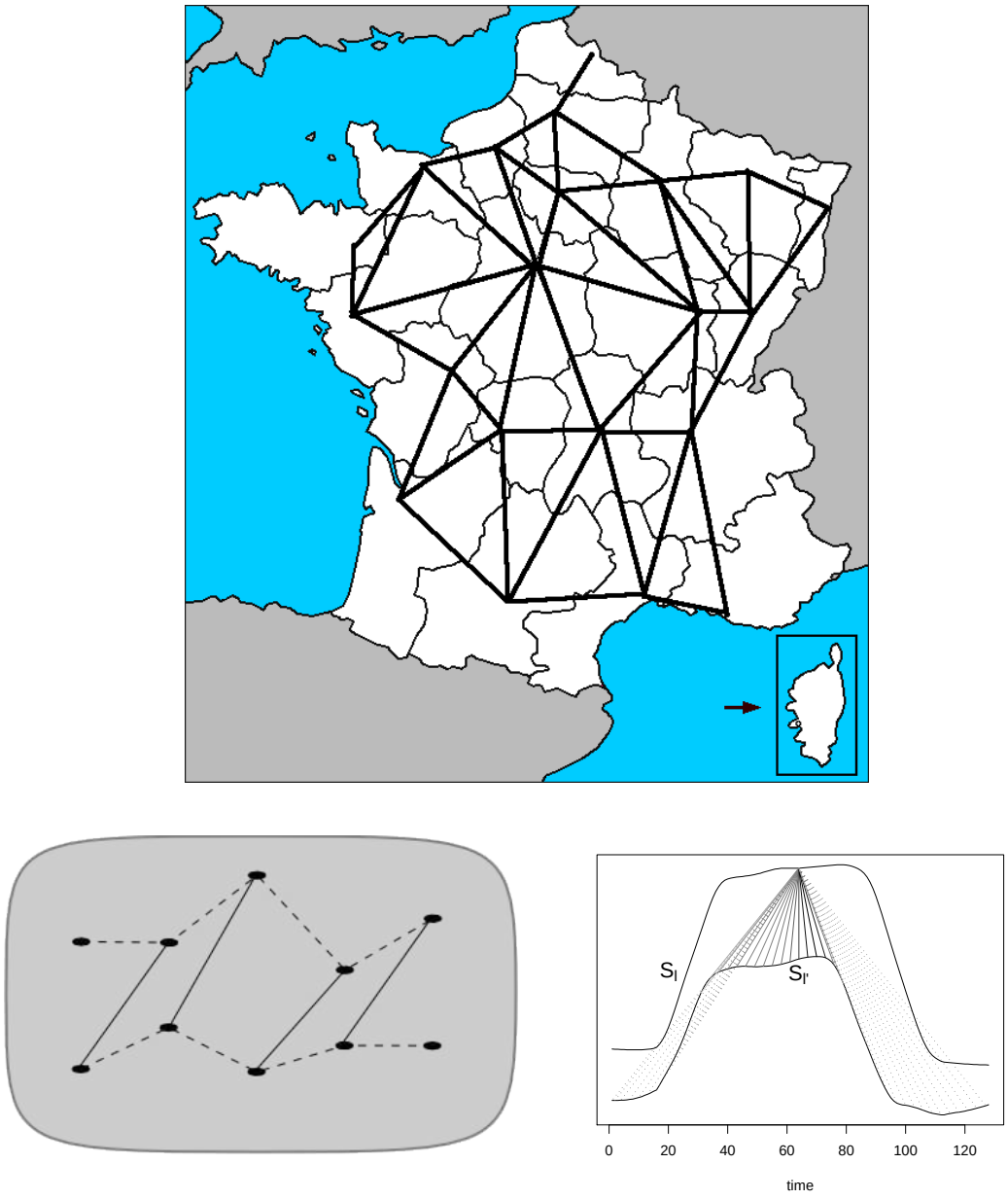


FIGURE 9 – Différents types de structures de contiguïté

(1969) propose de séparer les paires d'individus entre voisins et non voisins. Il calcule alors une matrice de variance-covariance "locale", i.e., limitée aux paires d'individus voisins. Thioulouse *et al.* (1995) proposent une autre matrice de covariance entre les voisins, inspirée des

travaux de Wartenberg (1985), nommée dans la littérature "globale", par opposition à la variance locale de Lebart. En pratique, ces deux approches visent à généraliser deux indices d'autocorrélation spatiale proposés par Geary (1954) et Moran (1950). L'objectif du début de ce chapitre (partie 1) est d'étudier ces indices pour cerner leurs spécificités. Ces indices permettent de définir plusieurs extensions de la notion classique de matrice de variance-covariance au cas de données contiguës (étudiées dans la partie 2), qui s'étendent au cas particulier d'ensembles et de partitions de séries temporelles (partie 3). En particulier, ces définitions permettent d'introduire des analyses exploratoires particulières, qui consistent à chercher les composantes principales au sens des projections révélatrices d'un critère.

1 Etude des indices d'autocorrélation spatiale usuels

Les notions de contiguïté temporelles et géographiques sont similaires. Les données recueillies au sein d'une population d'individus liés par de telles structures comportent souvent une dimension géographique ou temporelle. Par exemple, un paramètre étudié sur toutes les régions d'un pays peut présenter des comportements voisins au sein de régions proches géographiquement. De même, une proximité temporelle se traduit souvent par des observations proches. D'un point de vue temporel, s'il pleut à un moment dans une région, il y a un plus grand risque qu'il pleuve dans cette région une heure plus tard. D'un point de vue géographique d'autre part, le risque est plus important qu'il pleuve dans la région voisine. Cependant, dans d'autres cas, la proximité temporelle ou géographique peut être à l'origine de fortes différences. Par exemple, géographiquement, le découpage des régions administratives a été fait sur la base de différences réelles : différence de relief, climat, type de sol, dialecte... De plus, des différences apparaissent quant à la politique mise en œuvre... De ce fait, l'observation d'un indicateur sur les régions voisines peut mettre à jour des différences importantes. Un autre exemple concerne des séries temporelles : une forte baisse d'un indicateur économique peut être suivie d'un rebond haussier compensatoire.
De manière générale, il peut exister une forte opposition entre observations contiguës pour certains critères, et ces différences peuvent être importantes à considérer. Nous évaluons la ressemblance et l'opposition des régions contiguës par des indices d'autocorrélation spatiale. Deux indices spatiaux classiques sont les indices de Moran et de Geary.
Cette partie vise à étudier ces deux indices à travers leurs bornes et leur comportement.

1.1 Notation

Soit donnée une population de taille n. Nous considérons dans la suite un vecteur $x \in \mathbb{R}^n$, et $y \in \mathbb{R}^n$ le vecteur centré de x au sens de la métrique considérée. On note I_M l'indice de Moran défini par

$$I_M = \frac{n}{\sum_{ij} w_{ij}} \frac{\sum_{ij} w_{ij} y_i y_j}{\sum_j y_j^2}$$

On note I_C l'indice de Geary défini par

$$I_C = \frac{n-1}{2\sum_{ij} w_{ij}} \frac{\sum_{ij} w_{ij} (y_i - y_j)^2}{\sum_j y_j^2}$$

Si on suppose en sus que les poids sont normalisés, i.e., $\sum_{i,j=1}^{n} w_{ij} = 1$, les indices se réécrivent alors de la manière suivante :

$$I_M = \frac{\sum_{ij} w_{ij} y_i y_j}{\frac{1}{n}\sum_j y_j^2}$$
$$I_C = \frac{1}{2}\frac{\sum_{ij} w_{ij}(y_i - y_j)^2}{\frac{1}{n-1}\sum_j y_j^2}$$

Dans le cadre d'une réécriture matricielle, il est important d'introduire certaines notations supplémentaires.

Notations matricielles Soit $\mathbb{1}_n$ le vecteur $(1, \ldots, 1)$ et I_n la matrice identité.
Soit $X_{(n\times 1)} = [x_i]$, le vecteur décrivant l'observation de n individus pour la variable $\mathbf{X}$. x_i est l'observation de la variable X pour l'individu i.
Soit $N_{(n\times n)}$=diag($\frac{m_{1.}+m_{.1}}{2}$,...,$\frac{m_{n.}+m_{.n}}{2}$) la matrice diagonale des poids des instants ; N vérifie $\sum n_{ii} = 1$.
Soit $Y_{(n\times p)}=(I_n - \mathbb{1}_n^t \mathbb{1}_n N)X$ la matrice N centrée de X.

1.2 Bornes des indices

L'étude du domaine dans lequel ces indices évoluent est fondamentale. En particulier, à des fins de comparaison, il est important que l'indice d'autocorrélation soit borné. Dans le cas de voisinages équilibrés (même nombre de voisins pour chaque élément), les indices évoluent entre -1 et 1 pour Moran, et entre 0 et 2 pour Geary. En revanche, lorsque la taille de la population n'est pas fixe, les indices de Moran et de Geary ne sont pas bornés uniformément. Les bornes des indices dépendent de la taille de la population. Ainsi, dans certaines conditions très particulières, l'indice de Moran prend des valeurs comprises entre -n et n (n étant la taille de la population). C'est le cas de l'exemple dans la figure 10.
Notons que le fait de normaliser l'indice (division par n par exemple) n'est pas adéquat, car il privilégie des cas très rares au détriment de cas usuels.

De la même façon, l'indice de Geary est positif, mais n'admet pas une majoration uniforme (Figure 11).

Dans ces deux exemples, les problèmes sont similaires et apparaissent du fait d'un fort déséquilibre entre nombres de voisins pour chaque individu. Les individus qui ont le plus de voisins dominent les autres, et peuvent contribuer fortement à la covariance des voisins. C'est dû au fait de n'imposer une contrainte de poids de voisinage que globale. Nous allons voir que le fait d'imposer un poids de voisinage constant (e.g., $\frac{1}{n}$) pour chaque point, à la fois sur les arêtes entrantes et sortantes, contraint l'indice de Moran à rester entre -1 et 1, et l'indice de Geary entre 0 et 2.

1.3 Théorème

Les preuves de ces résultats sont placées dans l'annexe D

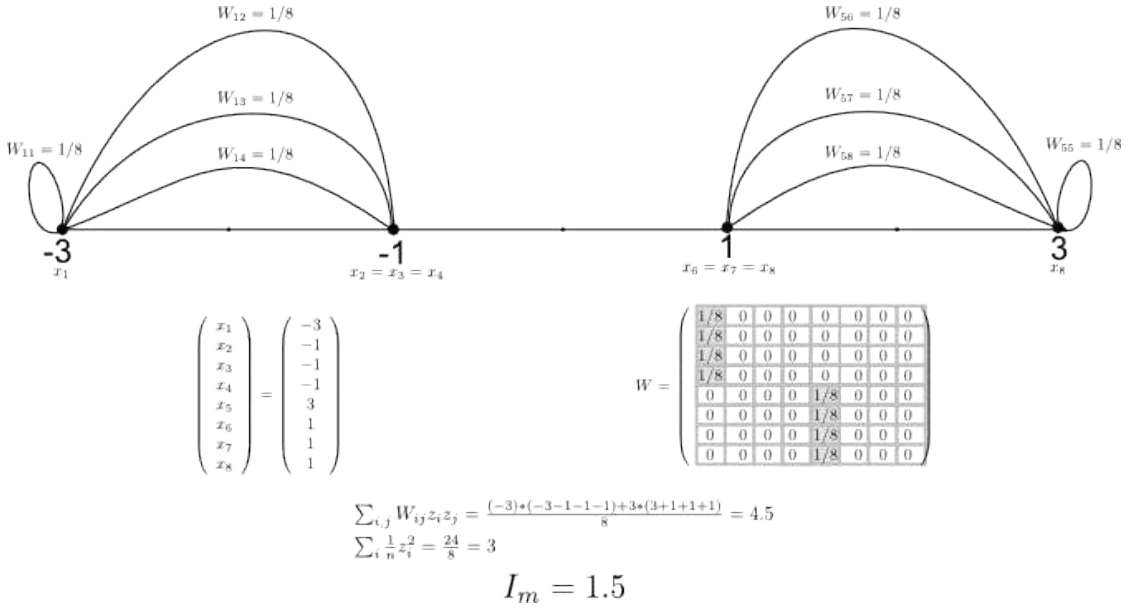


FIGURE 10 – Exemple où l'indice de Moran est plus grand que 1 quand la somme sur les colonnes n'est pas égale à 1

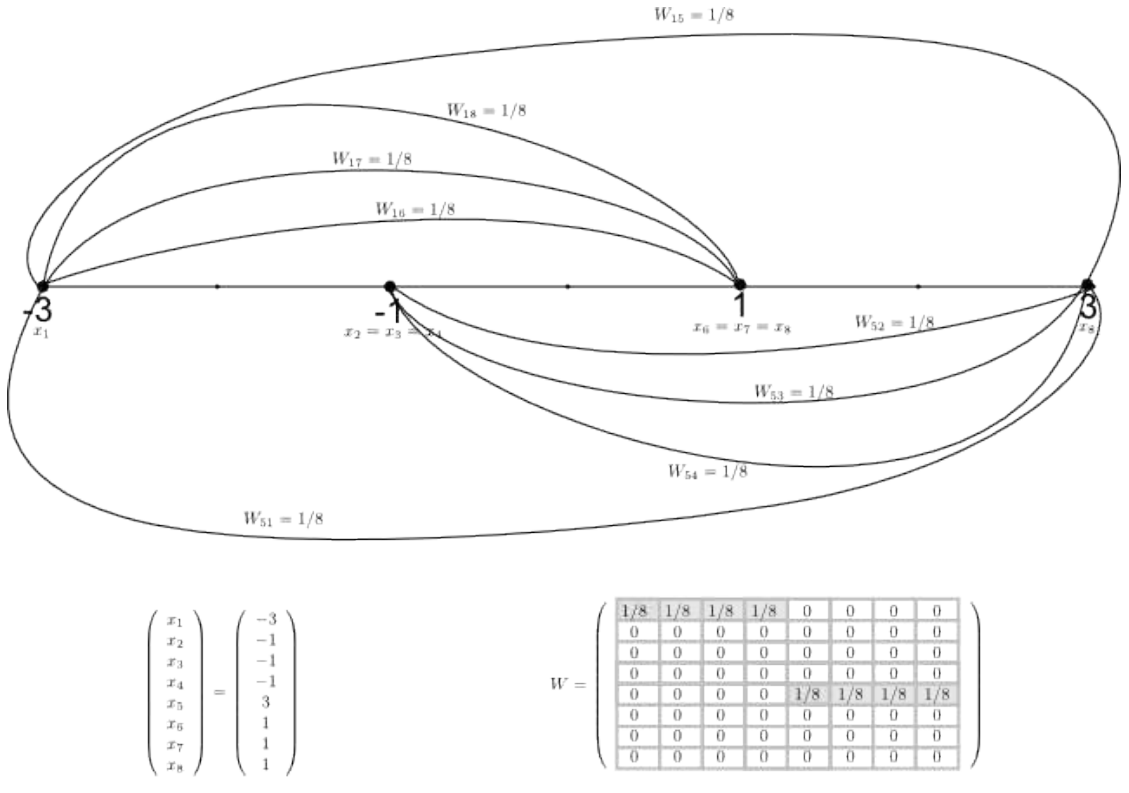


FIGURE 11 – Exemple où l'indice de Geary est plus grand que 2 quand la somme sur les lignes n'est pas égale à 1

Proposition 16 : (Borne de l'indice de Moran)

Si $\forall i_0, j_0 \in \{1, \ldots, n\} \sum_{i=1}^{n} w_{ij_0} = \sum_{j=1}^{n} w_{i_0 j} = \frac{1}{n}$, alors $I_M \in [-1, 1]$

Remarque 17 : (Un exemple d'application)

Dans le cas où les poids de voisinages sont symétriques et égaux à $\frac{1}{n}$ par ligne, nous sommes dans les conditions d'application du théorème précédent.

1.4 Relation entre les indices d'autocorrélation spatiale de Geary et de Moran

Plaçons-nous dans le cas précédent où les poids des voisinages en ligne et en colonnes sont égaux à $\frac{1}{n}$.

Proposition 18 :

$$Si\ \forall i_0, j_0 \in \{1, \ldots, n\} \sum_{j=1}^{n} w_{i_0 j} = \sum_{i=1}^{n} w_{i j_0} = \frac{1}{n},\ alors\ I_C = \frac{n-1}{n}(1 - I_M)$$

En particulier, I_C évolue dans $[0, 2 - \frac{2}{n}]$

1.5 De nouveaux indices

Les exemples précédents mettent en lumière le fait que ces indices reposent sur le prédicat suivant : chaque individu joue un même rôle dans le calcul de cet indice. Pour régler ces problèmes, Cliff et Ord (1972) proposent de modifier la métrique au dénominateur, en pondérant la variance par la métrique diagonale $N_{(n\times n)} = \frac{1}{2}(w_{i.} + w_{.i})_{i\in\{1..n\}}$.

Définition 19 : (Indice de Moran corrigé)
L'indice d'autorégression de Moran corrigé s'écrit

$$I_M* = \frac{\sum_{ij} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i N_{ii}(x_i - \bar{x})^2} \tag{10}$$

$$avec\ \bar{x} = \sum_i N_{ii} x_i \tag{11}$$

Définition 20 : (Indice de Geary corrigé)
L'indice d'autorégression de Geary corrigé s'écrit

$$I_C* = \frac{\sum_{ij} w_{ij}(x_i - x_j)}{\sum_i N_{ii}(x_i - \bar{x})^2} \tag{12}$$

$$avec\ \bar{x} = \sum_i N_{ii} x_i \tag{13}$$

Dans le cas de ces indices, le centrage se fait par la métrique N des poids de voisinage.

Proposition 21 : (Bornes des indices corrigés)

Avec cette modification des indices, l'indice de Geary varie entre 0 et 2, et l'indice de Moran varie entre -1 et 1. Les deux indices sont liés par la relation $I_C* = 1 - I_M*$

Nous utiliserons dans la suite les notations I_M et I_C pour décrire les indices corrigés.

1.5.a Ecriture matricielle des indices

Ces indices, définis à partir de leur écriture analytique, ont également une écriture matricielle.

Proposition 22 : (Indice de Moran corrigé)
L'indice de Moran s'écrit $I_M = \frac{{}^tYWY}{{}^tYNY}$

La preuve de ce résultat est immédiate en réécrivant le produit matriciel.

Proposition 23 : (Indice de Geary corrigé)

L'indice de Geary s'écrit $I_c = \frac{{}^tY(N-\tilde{W})Y}{{}^tYNY}$ *où* $\tilde{W} = \frac{W+{}^tW}{2}$

Remarque 24 : (Symétrie)

Nous remarquons avec l'écriture précédente que quelle que soit la matrice W, le numérateur de l'indice de Geary est toujours une forme quadratique, même si le voisinage n'est pas symétrique

1.6 Quelques valeurs particulières

Pour illustrer la façon dont les deux indices de Moran et de Geary fonctionnent, nous étudions trois cas de séries temporelles.

1. Le premier est le cas où les données voisines sont homogènes au sein d'un voisinage. Nous étudions le cas d'un découpage de l'intervalle de temps en 3, où le voisinage est constitué de cliques avec des valeurs égales sur chacune.
2. Le deuxième cas sera celui d'une répartition sans voisinage, c'est-à-dire un graphe de voisinage complet.

3. Le dernier cas est celui où les données entre voisins ont tendance à s'opposer, en observant une série temporelle selon le voisinage suivant : deux instants sont voisins s'ils correspondent à des instants successifs.

1.6.a Cas d'un voisinage homogène

Soit $P = \{P_1, P_2, P_3\}$ une partition de $x_1 ... x_T$.
On pose $W_{ij} = \frac{1}{n\#(P_k)}$, si x_i et x_j appartiennent à la même classe P_k, 0 sinon.
Soit z la variable centrée valant z_k sur la classe P_k.
L'indice de Geary vaut

$$I_C = \frac{\sum_{ij} w_{ij}(z_i - z_j)}{\frac{1}{n}\sum_j z_j^2} = \frac{\sum_{k=1}^{3} \sum_{j \in P_i} \frac{1}{\#(P_i)}(z_k - z_j)^2}{\sum_j z_j^2} = 0$$

Par la relation qui lie les indices, l'indice de Moran vaut 1.

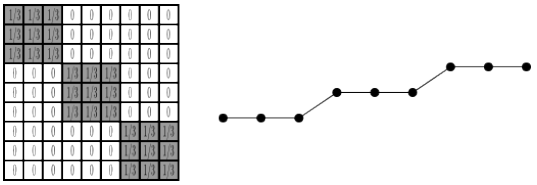


FIGURE 12 – Cas d'un voisinage homogène

1.6.b Cas d'un voisinage complet

Soit $W_{i,j} = \frac{1}{n^2}$. $N = \frac{1}{n} I_n$.
Soit z une variable N-centrée quelconque, i.e., $\forall j \in \{1, \ldots, n\}$, $z_j = -\sum_{i \neq j} z_i$

$$I_M = \frac{\sum_{ij} w_{ij} z_i z_j}{\sum_j N_{jj} z_j^2} = \frac{\sum_{ij} \frac{1}{n} z_i z_j}{\sum_j z_j^2}$$

$$I_M = \frac{\sum_i \frac{1}{n} z_i \sum_{j \neq i} z_j}{\sum_j z_j^2} + \frac{\sum_i \frac{1}{n} z_i^2}{\sum_j z_j^2}$$

$$I_M = \frac{\sum_i \frac{1}{n} - z_i^2}{\sum_j z_j^2} = 0$$

L'indice de Moran vaut 0. A nouveau par la relation qui lie les indices, l'indice de Geary vaut 1.

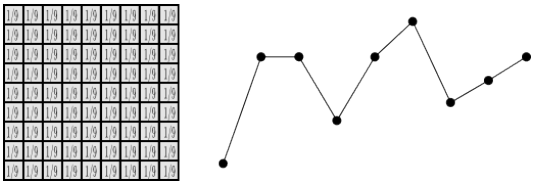


FIGURE 13 – Cas d'un voisinage complet

1.6.c Cas d'un voisinage hétérogène

Soit W un voisinage en chaîne ; on pose $W_{ij} = 1$ si $i = j + 1$.
$I_C = \frac{1}{n-1} \sum_i (z_i + 1 - z_i)^2$
$I_M = \frac{1}{n-1} \sum_i z_i z_{i+1}$
Remarquons que, dans ce cas, I_C est la covariance temporelle, et I_M est l'autocorrélation temporelle.

Soit z la variable $(-1)^n z_0$ (on suppose qu'on a un nombre pair de termes).
L'indice de Moran vaut

$$I_M = \frac{\sum_{ij} w_{ij} z_i z_j}{\sum_j z_j^2} = \frac{1}{n-1} \frac{\sum_{i=1}^{n} z_i z_{i+1}}{\sum_j N_{jj} z_j^2} = \frac{\sum_{j=1}^{n} -z_0^2}{\sum_j N_{jj} z_0^2} = -1$$

A nouveau, l'indice de Geary vaut 2.

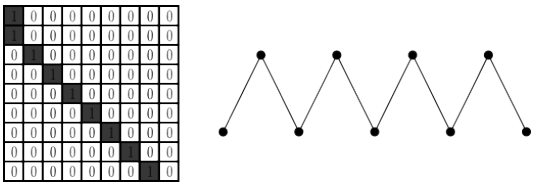


FIGURE 14 – Cas d'un voisinage hétérogène

1.7 Interprétation

Nous voyons sur les exemples précédents que l'indice de Geary est plus élevé pour des variables ayant des zones de forte hétérogénéité, le maximum étant atteint pour une structure où toutes valeurs voisines s'opposent (structure en échiquier). Il prend ses plus faibles valeurs pour les régions les plus homogènes, c'est-à-dire où les observations voisines prennent des valeurs très proches, le minimum étant atteint pour une variable constante. En effet, le numérateur de l'indice de Geary est une variance limitée aux individus voisins.
Au contraire, l'indice de Moran prend ses valeurs maximales lorsque les données voisines sont proches et ses valeurs minimales lorsque les voisins s'opposent. L'indice de Moran a un numérateur qui se présente en effet comme une covariance entre les individus voisins.
L'indice de Geary s'apparentant à une variance limitée aux individus voisins, nous allons, dans la partie suivante, définir une extension de la variance fondée sur ces indices d'autocorrélation spatiale, adaptée à une structure de contiguïté. Notons cependant que les valeurs des indices de Moran et de Geary dépendent de deux critères, d'une part la structure de contiguïté, et d'autre part, les valeurs prises par les différents individus.
Dans certaines configurations particulières, à l'instar des trois configurations décrites ci-dessus, tantôt le voisinage, tantôt les valeurs imposent une valeur fixée aux deux indices, rendant leur interprétation plus complexe. La comparaison n'est possible que si le voisinage est fixé a priori.

2 Variance associée à une structure de contiguïté

La variance est un résumé statistique d'une distribution permettant de caractériser la dispersion des valeurs par rapport à la moyenne, tandis que la covariance est une mesure de

variation simultanée, elle évalue le lien linéaire entre deux observations. Ces quantités sont classiquement calculées sur l'ensemble des individus (ici, les instants des séries temporelles). Nous souhaitons limiter le calcul de la matrice de variance-covariance aux individus voisins. Rappelons dans un premier temps la définition de la matrice usuelle de variance/covariance, avant de présenter sa généralisation aux données contiguës. Nous introduirons dans la partie suivante une formalisation dans le cadre des données issues de séries temporelles.
Nous notons $\mathbb{1}_n$ le vecteur $(1, \dots 1)$, I_n la matrice identité, et U_n la matrice unitaire $U_n = \mathbb{1}_n\ {}^t\mathbb{1}_n$.

$$U_n = \begin{array}{c} 1 \\ \vdots \\ n \end{array}\!\!\begin{pmatrix} 1 & \dots & 1 \\ 1 & \vdots & 1 \\ 1 & \dots & 1 \end{pmatrix} \qquad I_n = \begin{array}{c} 1 \\ \vdots \\ n \end{array}\!\!\begin{pmatrix} 1 & 0 & \\ 0 & \ddots & 0 \\ & 0 & 1 \end{pmatrix}$$

Nous appelons X la matrice de dimension $(n \times p)$ décrivant n observations selon p variables numériques $\mathbf{X}_1, \dots \mathbf{X}_p$.. $X_{(n\times p)} = [x_{ij}]$ avec x_{ij} l'observation de la variable j pour l'individu i.

$$X = \begin{array}{c} \\ n_1 \\ \vdots \\ n_n \end{array}\!\!\begin{array}{c} \begin{array}{cccc} X_1 & \dots & \dots & X_p \end{array} \\ \begin{pmatrix} x_{11} & \dots & \dots & x_{1p} \\ \vdots & \vdots & x_{i,j} & \vdots \\ x_{n1} & \dots & \dots & x_{np} \end{pmatrix} \end{array}$$

Soit $N_{(n\times n)}$ la matrice diagonale des poids des instants, de terme général $n_{ii} = \frac{m_{k.}+m_{.k}}{2}$.
Les poids des instants vérifient la relation suivante : $\sum n_{ii} = 1$
$N_{(n\times n)}$=diag($\frac{m_{1.}+m_{.1}}{2}, \dots, \frac{m_{n.}+m_{.n}}{2}$).
Soit $Y_{(n\times p)}$=$(I_n - \mathbb{1}_n^t \mathbb{1}_n N)X$ la matrice N centrée de X.
La matrice de variance/covariance usuelle est la matrice suivante :

$$V = X^t(I_n - U_nP)^tP(I_n - U_nP)X \tag{14}$$

où, I_n est la matrice diagonale Identité, U_n la matrice Unitaire, et P une matrice diagonale de poids dont le terme général p_i est égal à $\frac{1}{n}$, dans le cas particulier où toutes les observations sont équipondérées.

Redéfinissons la matrice de variance/covariance dans le cadre d'une structure de contiguïté. Nous utilisons pour cela les indices d'autocorrélation spatiale rencontrés dans les paragraphes précédents.
L'idée de départ de Lebart est de définir une covariance limitée aux individus voisins, qui étend l'indice de Geary à un ensemble de données multivariées. Wartenberg (1985) prolonge cette généralisation à l'indice de Moran. Ces approches, décrites dans un contexte spatial, se généralisent à un contexte temporel sans difficulté.

Reprenons l'écriture matricielle précédente.

2.1 Variance locale : fondée sur l'indice de Geary

Proposée par Lebart (1969), cette approche généralise l'indice de Geary. Elle repose sur la réécriture de la variance comme somme sur tous les couples d'observations :
Notons $\bar{x} = \sum_{i=1}^{n} p_i x_i$. Si P décrit un système de poids, (i.e. $\sum_{i=1}^{n} p_i = 1$), alors :

$$\sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j (x_i - x_j)^2 = 2\sum_{i=1}^{n} p_i (x_i - \bar{x})^2 \tag{15}$$

(preuve en annexe) Nous pouvons alors décomposer cette formule en deux sommes, la première sommant sur tous les couples d'observations voisines, et la seconde sur toutes les observations non voisines.

Définition 25 : (Variance locale)
En renormalisant la somme précédente portant sur les arêtes voisines, la variance locale introduite par Lebart est

$$V_l(x) = \frac{1}{2m}\sum_{i=1}^{n} \sum_{i' voisin\ de\ i} (x_i - x_{i'})^2$$

avec m le nombre de voisins, ce qui se généralise de la manière suivante, sous certaines conditions de normalisation de la matrice $N - W$ par la matrice V_l définie par

$$V_l = {}^t Y(N - W)Y$$

Nous avons alors

$$V_l(x) = \frac{1}{2\sum_{i=1}^{n}\sum_{i'=1}^{n} m_{ii'}} \sum_{i=1}^{n}\sum_{i'=1}^{n} m_{ii'}(x_i - x_{i'})^2$$

Etudions quelques propriétés de la variance locale de Lebart.

Remarque 26 : (symétrie)
Dans le cas d'une matrice W symétrique, la matrice $N - W = N - \tilde{W}$ est la matrice d'une forme quadratique semi-définie positive. La variance locale se présente comme une variance classique limitée aux paires d'observations voisines.

2.2 Variance globale : fondée sur l'indice de Moran

Proposée par Thioulouse *et al.* (1995), cette approche généralise l'indice de Moran.

Définition 27 : (Variance globale)
La variance globale est la matrice V_g définie par

$$V_g = {}^t YWY$$

$$V_g(x) = \frac{1}{2\sum_{i=1}^{n}\sum_{i'=1}^{n} m_{ii'}} \sum_{i=1}^{n}\sum_{i'=1}^{n} m_{ii'}(y_i - y_{i'})^2$$

Le terme de variance globale provient d'un découpage de la variance totale en deux parties, la variance locale (définie au sens de Lebart), et la covariance locale. La variance se décompose donc en deux composantes, la première étant dite locale, la seconde est appelée globale par abus de langage.

Remarque 28 :
A la différence de la variance locale précédente, la variance globale n'est pas une forme quadratique. En particulier, elle peut admettre des valeurs propres négatives.

Le nom de variance globale est utilisé en opposition à la variance locale de Lebart. Cette quantité s'approche d'une covariance entre les données X et les données centrées WX. De par son caractère non définie positive, cette expression est délicate à étudier en analyse multivariée, du fait de la présence possible de valeurs négatives, mais donne dans de nombreuses situations des résultats pratiques intéressants.

Remarque 29 : (Lien avec l'indice de Moran)
A partir de cette définition, on remarque que les p termes diagonaux de cette matrice correspondent aux indices de Moran des p variables.

Ces formules se généralisent aisément au cas de matrices de contiguïté pondérées.

2.3 Un nouveau formalisme introduit par Mom (1988)

L'objectif de Mom (1988) dans sa thèse est de généraliser l'analyse discriminante au cas où les individus ne sont pas regroupés par classe mais sont plus généralement reliés par une relation de voisinage. Il considère la définition sans biais de la variance locale de Lebart.

$$V_L = \frac{n}{2(n-1)K_1} \sum_{i=1}^{n} \sum_{j \in V_i} (x_i - x_j)^2$$

Une modification est faite sur la définition précédente pour rendre la variance locale compatible avec la variance intra quand la structure de contiguïté correspond à un découpage en partition des observations (i.e. une série de cliques séparées). La variance locale se redéfinit comme :

$$V(M) = \frac{1}{n} \sum_{i=1}^{n} (y_i - m(M)_i)^2 \tag{16}$$

où $m(M)_i$ est le barycentre des observations $x_{i'}$ voisines de i, $m(M)_i = \frac{1}{\sum_{i'} m_{ii'}} \sum_{i'} m_{ii'} x_{i'}$. En notant V_{L_1} la matrice de covariance locale et V_{L_2} la matrice de covariance locale sur le graphe de non-contiguïté, la matrice de covariance totale est donc somme des matrices V_{L_1} et V_{L_2}. A l'image de l'analyse discriminante avec les matrices intra et inter W et B, on étudie les valeurs propres et les vecteurs propres associés à la matrice $V_T^{-1} V_{L_2}$. De par la relation précédente, nous pouvons, comme pour l'analyse discriminante, relier ces vecteurs propres aux vecteurs propres de $V_{L_1}^{-1} V_{L_2}$, dès lors que V_{L_1} est inversible.

2.4 Analyse de contiguïté

Les méthodes factorielles telles que l'analyse en composantes principales sont fondées sur la diagonalisation des matrices de variance-covariance et découlent du théorème d'inertie.

La matrice de variance-covariance est symétrique et ses valeurs propres sont positives. Le théorème nous assure que l'inertie est maximale pour les vecteurs propres associés aux plus fortes valeurs propres. Les matrices de variance-covariance globale $V_g = {}^tZ \times W \times Z$ et locale $V_l = {}^t Z(N-W)Z$ ne sont plus symétriques en général, et rien n'assure la positivité des valeurs propres. Cependant, les valeurs propres existent et sont réelles, et il reste possible de s'intéresser aux valeurs propres maximales.

2.4.a Analyse locale

La méthode proposée par Lebart (1969) correspond à l'analyse du tableau X de taille n, p. On note $N = diag[w_1, ..., w_n]$ la matrice diagonale des poids de voisinage, où $w_i = \frac{1}{2}\sum_j (w_{ij} + w_{ji})$, $V_L = [v^I_{jk}]$ la matrice de variance/covariance tenant compte de la structure de contiguïté spatiale, et $R_L = [r^I_{jk}]$ la matrice de corrélation spatiale fondée sur l'indice de Geary. A l'instar d'une analyse en composantes principales, l'objectif est de diagonaliser ces deux matrices.

$$V_L = Y^t(N-W)Y$$
$$R_L = Z^t(N-W)Z$$

$$\text{avec } v^L_{jk} = Y_k = \frac{1}{2}\sum_{i,i'} w_{ii'}(y_{ij} - y_{i'k})^2$$
$$\text{et } r^L_{jk} = \frac{1}{2}\frac{\sum_{i,i'} w_{ii'}(y_{ij} - y_{i'j})(y_{i'j} - y_{i'k})}{\sqrt{\frac{1}{n}\sum_i y^2_{ij}}\sqrt{\frac{1}{n}\sum_i y^2_{ik}}}$$

La matrice R_L porte sur sa diagonale les indices de Geary de chaque variable. Les vecteurs propres associés aux plus grandes valeurs propres de la matrice de variance/covariance V_L sont ceux pour lesquels l'homogénéité est maximale au cœur des voisinages.

2.4.b Analyse globale

L'analyse en composantes principales proposée par Thioulouse *et al.* (1995) est l'analyse du tableau Y de taille n, p, avec Y la matrice X centrée en ligne, $Y = (I - 1_n \times {}^t 1_n \times D) \times X$, 1_n le vecteur de dimension $(1 \times n)$ constitué de 1, et D la matrice des poids des observations. Elle consiste à rechercher les valeurs propres de la matrice de variance/covariance globale $V_G = {}^t Y \times W \times Y$.
On note $V_G = [v^I_{jk}]$ et $R_G = [r^I_{jk}]$ les matrices de variance/covariance et de corrélation qui incluent la structure de contiguïté spatiale.

$$V_G = Y^t W Y$$
$$R_G = Z^t W Z$$

$$\text{avec } v_{jk}^G = \sum_{i,i'} w_{ii'} y_{ij} y_{i'k}$$

$$\text{et } r_{jk}^G = \frac{\sum_{i,i'} w_{ii'} y_{ij} y_{i'k}}{\sqrt{\frac{1}{n}\sum_i y_{ij}^2}\sqrt{\frac{1}{n}\sum_i y_{ik}^2}}$$

Les termes diagonaux de la matrice R_g sont les indices de Moran des variables X_j. Le vecteur propre associé à la plus grande valeur propre correspond à la combinaison linéaire des variables maximisant l'indice de Moran. Cela correspond au vecteur dont les voisinages sont les plus homogènes.

La matrice R_g n'est pas semi-définie positive a priori. Ceci est une différence importante par rapport à la matrice de corrélation classique. Elle est cependant toujours diagonalisable réelle. A présent, cependant, certaines valeurs propres peuvent être négatives. Les valeurs propres négatives ont également un sens. Elles correspondent aux vecteurs les plus hétérogènes au cœur de leur voisinage, tandis que les valeurs propres positives correspondent à un voisinage homogène.

2.4.c Analyse de Mom

L'analyse classique, qui consiste à éloigner les centres de gravité des séries et à rapprocher tous les points au sein d'une même classe autour du centre de gravité ne tient pas compte de la structure particulière des séries temporelles au niveau des courbes. En vue de notre travail sur les séries temporelles, nous aimerions étendre le cadre discriminant en utilisant la notion de contiguïté.

Les travaux de Mom (1988) dans sa thèse généralisent l'analyse discriminante à une structure de contiguïté. Il définit une métrique qui ne limite plus l'analyse à la recherche de directions compatibles avec une répartition des données en classes, mais plus généralement, il recherche des directions adaptées à la structure de graphe décrivant le voisinage. C'est une généralisation de l'analyse factorielle discriminante, où la répartition en classes est le cas particulier d'un graphe de voisinage constitué de cliques.

Il s'inspire de l'analyse de contiguïté de Lebart et décompose la variance totale $V_T = X' \frac{nI_n - U}{n^2} X$ en fonction de V_L et $V_{L'}$ avec $V_L = \frac{{}^tX[(N-W)]X}{K}$ et $V_{L'} = \frac{{}^tX[(N'-W')]X}{K'}$, où W est la matrice booléenne associée au graphe (ou matrice d'incidence), N la matrice diagonale où le $j^{\text{ième}}$ élément est le nombre de voisins du sommet j, et K le nombre total d'arêtes du graphe. N', W' et K' sont les correspondants pour le graphe complémentaire. Mom obtient alors la relation

$$V_T = \frac{K}{K+K'} V_L + \frac{K'}{K+K'} V_L'$$

qui assure une équivalence entre les vecteurs propres de $\frac{V_L}{V_T}$ et de $\frac{V_L}{V_{L'}}$. Dans sa thèse, Mom propose une métrique adaptée de la forme

$$M = N^{-1}(N-W)'D(N-W)N^{-1}$$

Critique de cette analyse Si le sens donné à la variance intra est assez claire, en ce qu'il s'agit de la matrice des différences locales, le sens donné à la variance inter est moins clair. Une autre critique à évoquer vient du fait que, à nouveau, tout repose sur un type de relation

de voisinage. On ne distingue pas le lien temporel du lien d'appartenance à une classe. Enfin, l'approche de Mom repose sur la matrice d'adjacence du graphe, et ne tient pas compte de la pondération des arêtes.

Extension à l'analyse discriminante L'analyse factorielle discriminante (AFD) est une méthode statistique fréquemment utilisée pour des tâches de classification supervisée. L'idée qu'il y a derrière l'AFD est celle de rapprocher les individus d'une même classe et de séparer au mieux les différentes classes. L'approche de Mom généralise l'AFD en cherchant à rapprocher les individus voisins et à éloigner les individus non-voisins. Il considère le fait d'être non-voisins comme le fait de ne pas avoir d'arête entre les deux points dans le graphe. Nous proposons une nouvelle relation de voisinage consistant à définir à la fois les voisins, et les non-voisins.

Pour cela, nous développons l'idée de Lebart de décomposer la variance totale en deux variances, mais ici, nous proposons de modifier à la fois la métrique de Lebart pour la variance intra, et la métrique pour la variance totale.
Soit W_1 et W_2 deux matrices d'incidence. Nous cherchons les axes qui rapprochent les sommets reliés par les arêtes de W_1 et éloignent les sommets reliés dans W_2. Notons V_{L_1} et V_{L_2} respectivement les variances fondées sur l'indice de Geary pour les voisinages W_1 et W_2. Nous cherchons à maximiser le rapport $\frac{V_{L_1}}{V_{L_1}+V_{L_2}}$. Nous allons pour cela nous assurer du fait que la matrice $V_{L_1} + V_{L_2}$ est inversible.
Posons $\tilde{N} = N_1 + N_2$ et $\tilde{W} = W_1 + W_2$. Naturellement, $\tilde{N} + \tilde{W} = \tilde{N}(I + \tilde{N}^{-1}\tilde{W})$. Dans le cas où la matrice $\tilde{W}$ est symétrique, la forme quadratique associée à $N - W$ est positive, du fait de la positivité de l'indice de Geary. Cependant, elle n'est pas forcément définie positive. En effet, dans le cas où la matrice de voisinage est constituée de sous-graphe disjoints, un vecteur constant sur chaque sous-graphe annule la forme quadratique. En pratique, un tel cas se produit avec une probabilité nulle. De façon générale, quand n est plus grand que p, la matrice ${}^tY(\tilde{W} + \tilde{N})Y$ est inversible (presque sûrement). Remarquons qu'au contraire de l'analyse discriminante, si la matrice n'est pas inversible, on ne peut pas se contenter de faire une ACP au préalable, car des données décorrelées au sens de la métrique D ne le seront pas forcément au sens de la métrique $N - W$.

Choix des matrices de voisinage en vue de l'analyse Dans ce qui précède, nous avons pu observer l'importance du choix des matrices de contiguïté. Nous présentons dans l'annexe E de ce document une étude plus détaillée sur le choix des matrices de voisinage. Chaque définition a priori d'une matrice de voisinage peut conduire à une nouvelle analyse, au sens d'une des trois approches. Nous nous sommes penchés sur les analyses suivantes.

Les résultats associés à chacune de ces analyses, présentés en annexe, montrent la différence fondamentale qu'induit le choix des diverses matrices et illustrent l'importance du choix des matrices de variance-covariance. Nous présentons dans la suite la généralisation des approches précédentes dans le cas de séries temporelles.

AFD Name	W	Fonction objectif	Type d'approche
AFD1	$W = (O, I, O)$	$max(\frac{V_I(W)}{V_T})$	Variance globale
AFD2	$W1 = (J, I, O)$	$max(\frac{V_I(W1)}{V_I(W1)+V_I(W2)})$	
	$W_2 = (O, O, I)$		
AFD3	$W = (O, O, U)$	$max(\frac{V_c(W)}{V_T})$	Variance locale
AFD4	$W_1 = (I, O, I)$	$max(\frac{V_c(W_1)}{V_c(W_1)+V_c(W_2)})$	
	$W_2 = (O, I, O)$		
AFD5	$W = (O, U, O)$	$max(\frac{VB(W)}{V_T})$	Variance de Mom

TABLE 1 – Approches discriminantes

3 Variance induite par des séries temporelles

Les objets qui nous intéressent sont des séries temporelles. Nous employons pour cela le formalisme introduit dans la section précédente dans le cadre de données contiguës, en l'étendant aux ensembles de séries temporelles. Nous introduisons ensuite une structure supplémentaire de partition des séries temporelles en classes. Rappelons les notations définies a chapitre 1.

Notations On note une série temporelle par une matrice de dimension $T \times p$ où T est le nombre d'instants qui caractérisent la série et p le nombre de variables.

$$X = \begin{array}{c} \\ t_1 \\ \vdots \\ t_T \end{array} \begin{array}{c} \begin{array}{ccc} X_1 & \dots & X_p \end{array} \\ \begin{pmatrix} S^1_{11} & \dots & S^1_{1p} \\ \vdots & S^1_{ij} & \vdots \\ S^1_{T1} & \dots & S^1_{Tp} \end{pmatrix} \end{array}$$

Dans le cas de plusieurs séries S^1,... S^n, on considère que les n séries sont exprimées sur chaque variable et ont même longueur T. On exprime les données par une matrice à p colonnes où sont concaténées les lignes des différents tableaux.

$$X = \begin{array}{c} \\ S^1 \\ \\ \\ \vdots \\ \vdots \\ \vdots \\ \\ S^n \\ \\ \end{array} \begin{array}{c} \begin{array}{ccc} X_1 & \dots & X_p \end{array} \\ \left(\begin{array}{ccc} S^1_{11} & \dots & S^1_{1p} \\ \vdots & S^1_{ij} & \vdots \\ S^1_{T1} & \dots & S^1_{Tp} \\ \hline & \vdots & \\ & \vdots & \\ & \vdots & \\ \hline S^n_{11} & \dots & S^n_{1p} \\ \vdots & S^1_{ij} & \vdots \\ S^n_{T1} & \dots & S^n_{Tp} \end{array}\right) \end{array}$$

3.1 Cas d'un ensemble de séries temporelles

Soit X la matrice $nT \times p$, décrivant n séries temporelles multivariées S^1, ... S^n à partir de p variables numériques observées au cours de T instants. L'appariement est décrit par une matrice M constituée de n^2 blocs matriciels de taille $T \times T$ notés $M^{(l_1,l_2)}$ avec $l \in \{1, ..., n\}$. Les blocs $M^{(l_1,l_2)}$ décrivent l'appariement entre la série S^{l_1} et la série S^{l_2}, son terme général $m^{l_1 l_2}_{i_1,i_2}$ caractérisant le poids du lien entre l'observation i_1 de la série S^{l_1} et l'observation i_2 de la série S^{l_2}.

Nous introduisons quatre types classiques d'appariements :

– Le couplage complet consiste à connecter toutes les observations de S^{l_1} et de S^{l_2}, indépendamment des instants. On note cette matrice, dite matrice unité, U_T.

$$U_T = \begin{array}{c} 1 \\ \vdots \\ T \end{array} \begin{pmatrix} 1 & \dots & 1 \\ 1 & \vdots & 1 \\ 1 & \dots & 1 \end{pmatrix} \tag{17}$$

– Le couplage Euclidien consiste à connecter toutes les observations de S^{l_1} et de S^{l_2}, apparaissant aux mêmes instants. On note cette matrice, dite matrice identité, I_T.

$$I = \begin{array}{c} 1 \\ \vdots \\ T \end{array} \begin{pmatrix} 1 & 0 & \\ 0 & \ddots & 0 \\ & 0 & 1 \end{pmatrix} \tag{18}$$

– Le couplage temporel consiste à connecter toutes les observations entre instants consécutifs. On note cette matrice, dite matrice temporelle, J_T.

$$J = \begin{array}{c} 1 \\ 2 \\ \vdots \\ T \end{array} \begin{pmatrix} 0 & 1 & 0 & \dots \\ 0 & 0 & \ddots & 0 \\ 0 & \ddots & 0 & 1 \\ \dots & 0 & 0 & 0 \end{pmatrix} \tag{19}$$

– Le couplage de type DTW est obtenu en cherchant au sein de l'ensemble des chemins possibles, l'alignement des séries minimisant leur écart cumulé. A la différence des autres appariements, le couplage DTW dépend des séries appariées.

$$M^{l_1,l_2} = \begin{array}{c} 1 \\ 2 \\ \vdots \\ T \end{array} \begin{pmatrix} 1 & 1 & 0 & \dots \\ 0 & 1 & 0 & \dots \\ \vdots & & \ddots & \\ 0 & 0 & \dots & 1 \end{pmatrix} \tag{20}$$

A partir de la matrice de voisinage, nous définissons la variance selon la formule de Mom.

Définition 30 : (Variance/covariance d'un ensemble de séries temporelles)
La matrice de variance/covariance V_M de dimension $(p \times p)$ induite par un ensemble de n séries $S^1, ..., S^n$ connectées selon la matrice d'appariement M se définit de la manière suivante.

$$V_M = X^t(I - M)^t P(I - M)X \tag{21}$$

avec P la matrice de poids diagonale $(nT \times nT)$, où $p_i = \frac{1}{n\,T}$ dans le cas de données équipondérées.

Dans un contexte univarié, la variance se réécrit :

Remarque 31 : (Variance dans le cadre univarié d'un ensemble de séries temporelles)
La variance V_M d'une variable X est donnée par la formule :

$$V_M = \sum_{l=1}^{n}\sum_{i=1}^{T} p_i \left(x_i^l - \sum_{l'=1}^{n}\sum_{i'=1}^{T} m_{i\,i'}^{l\,l'} x_{i'}^{l'} \right)^2 \tag{22}$$

Les valeurs x_{ij}^l sont centrées par rapport au terme $\sum_{l'=1}^{n}\sum_{i'=1}^{T} m_{i\,i'}^{l\,l'} x_{i'j}^{l'}$ qui évalue la moyenne des valeurs de X_j prises dans le voisinage de l'instant i de la série S^l. Le voisinage de i correspond à l'ensemble des instants i' de $S^{l'}$ $(l' \in \{1, ..., n\})$ connectés à i avec des poids $m_{ii'}^{ll'} \neq 0$.

3.2 Cas d'une partition de séries temporelles

Nous considérons à présent un contexte discriminant. L'objectif de ce paragraphe est d'étendre la définition précédente, décrivant la variance d'un ensemble de séries temporelles, au cas où l'ensemble des séries est partitionné en classes.
Considérons les séries $S^1, ..., S^n$ réparties au sein de K classes. On note $cl_l \in \{1, ..., K\}$ l'indice de la classe de la série S^l et n_k le nombre de séries temporelles constituant la classe k. Notons que $\sum_{k=1}^{n} n_k = n$.

La définition de la variance intra pour un ensemble de k classes de séries temporelles s'obtient à partir de l'expression définie au paragraphe précédent Eq.(21), sur la base d'un appariement intra-classe W.

Définition 32 : (Variance intra d'une partition de séries temporelles)
La variance intra, obtenue à partir d'une matrice d'appariement intra-classe W, prend la forme suivante :

$$WV_M = \frac{1}{nT}\sum_{k=1}^{K}\sum_{l=1}^{n_k}\sum_{i=1}^{T}(x_i^l - \frac{1}{n_k}\sum_{l'=1}^{n_k}\sum_{i'=1}^{T} m_{ii'}^{ll'}x_{i'}^{l'})^2$$

avec

$$M^{ll'} = \begin{cases} \mathbf{I} & si\ l = l' \\ \neq \mathbf{0} & si\ y_l = y_{l'}\ et\ l \neq l' \\ \mathbf{0} & si\ y_l \neq y_{l'} \end{cases} \quad (23)$$

où $\mathbf{I}$ *et* $\mathbf{0}$ *sont respectivement la matrice Identité et la matrice nulle de dimension* $(T \times T)$.

Cette caractérisation générale des blocs $W^{ll'}$ dans le contexte des matrices d'appariements intra W est dictée par trois types de contraintes.

- La première contrainte consiste à imposer une matrice Identité aux blocs diagonaux. Ceci implique un alignement euclidien entre chaque série et elle-même. En particulier, cela induit, pour chaque série comparée à elle-même, une variance nulle.
- La seconde contrainte oblige chaque série au sein d'une classe, à être liée à toutes les autres séries de la classe, i.e., pour chaque paire de série au sein d'une classe, il existe un couple d'instants liés Formellement, $\forall(l,l')/cl_l = cl_{l'}, \exists i,i' \in [1..T]/w_{ii'}^{ll'} \neq 0$. La variance intra doit être calculée sur la base de toutes les séries de la classe.
- Finalement, la troisième contrainte impose à chaque paire de séries de deux classes différentes de ne pas être liée, ces paires ne contribuant pas à la variance intra.

De manière similaire, la définition de la variance inter (correspondant à la variance entre les classes) induite par k classes de séries temporelles est obtenue avec la même définition de la variance, fondée à présent sur un appariement inter B restreignant le calcul de la variance aux séries temporelles appariées selon les règles ci-dessous.

Définition 33 : (Variance inter d'une partition de séries temporelles)
La variance inter, obtenue à partir d'une matrice d'appariement inter-classes B, prend la forme suivante :

$$\begin{aligned} BV_M = \quad & \tfrac{1}{nT} \sum_{k=1}^{K} \sum_{l=1}^{n_k} \sum_{i=1}^{T} (x_i^l \\ & - A_k (x_i^l + \sum_{k' \neq k} \sum_{l'=1}^{n_{k'}} \sum_{i'=1}^{T} m_{ii'}^{ll'} x_{i'}^{l'}))^2 \end{aligned}$$

avec

$$A_k = \frac{1}{1 + \sum_{k' \neq k, 1 \leq k, k' \leq K} n_{k'}}$$

et

$$M^{ll'} = \begin{cases} \mathbf{I} & si\ l = l' \\ \mathbf{0} & si\ y_l = y_{l'}\ et\ l \neq l' \\ \neq \mathbf{0} & si\ y_l \neq y_{l'} \end{cases} \tag{24}$$

où $\mathbf{I}$ *et* $\mathbf{0}$ *sont respectivement la matrice Identité et la matrice nulle de dimension* $(T \times T)$.

Cette caractérisation générale des blocs $B^{ll'}$ dans le contexte des matrices d'appariements inter B est le symétrique de la définition précédente dans le cadre des alignements intra. Elle est à nouveau dictée par trois types de contraintes très similaires aux contraintes intra-classe.

- La première contrainte consiste à imposer à nouveau une matrice Identité aux blocs diagonaux.
- La seconde contrainte oblige chaque série au sein d'une classe, à ne pas être liée aux autres séries de la classe,
- La dernière contrainte impose un couplage de toutes les paires de séries dans des classes différentes.

Remarque 34 : (Non-complémentarité des voisinages intra et inter)
Dans la définition précédente, il est important de noter que les voisinages intra et inter sont définis indépendamment. Mis à part les blocs diagonaux, les appariements intra et inter ne partagent aucune arête. Cependant, les structures intra et inter ne sont plus complémentaires ; certaines arêtes ont un poids nul dans les deux matrices d'appariement.

Conclusion

Nous avons présenté dans ce chapitre un formalisme adapté pour étendre les mesures utilisées dans le cadre de données contiguës à un ensemble ou à une partition de séries temporelles. Il apparaît nettement dans ce qui précède que l'appariement de séries temporelles joue un rôle crucial dans la définition des variances intra et inter. Un problème rencontré dans chaque approche réside dans le fait que la structure de contiguïté est toujours définie de manière ad hoc. Le lien temporel est introduit artificiellement. Le point fondamental au cœur de la discrimination de séries temporelles est alors la définition de telles matrices

d'appariement, sous les contraintes définies au sein des équations 23 et 24. Ce qui ressort de ces approches est la nécessité de définir des voisinages rapprochant les séries au sein des classes et éloignant les séries entre les classes pour la structure de voisinage associée à un ensemble de séries temporelles. Dans la suite, nous définirons une approche visant à apprendre des matrices d'appariements optimales, afin de minimiser la variance intra et maximiser la variance inter.

Conclusion de la partie I

Cette première partie présente un état de l'art des méthodes classiques pour la discrimination et la classification de séries temporelles. Il ressort de cette partie l'importance de bien apparier les instants des séries. Adopter une approche paire à paire nuit à la classification et il est préférable d'apparier simultanément de multiples séries temporelles. La notion de variance intra et inter-classes est un paradigme classique en analyse discriminante que nous souhaitons étendre au cadre temporel. Nous définissons alors des métriques fondées sur ces deux variances. La définition de ces métriques repose sur l'appariement des séries. Nous proposons alors dans la suite d'apprendre des appariements discriminants, au sens de la minimisation de la variance au sein des classes et de la maximisation de la variance entre les classes.

Partie II

Apprentissage des appariements

La première partie a mis en lumière l'importance du choix des appariements pour la définition d'une métrique adaptée à une partition de séries temporelles. Nous proposons, dans cette partie, d'apprendre des appariements discriminants. Pour cela, nous recherchons des appariements minimisant la variance intra et maximisant la variance inter-classes. La recherche de ces appariements se formalise comme un problème d'optimisation. Nous introduisons, dans le premier chapitre, deux algorithmes pour l'apprentissage des appariements intra et inter-classes. L'idée des algorithmes est de construire les appariements par pénalisation des poids des arêtes, en fonction de leur contribution à la variance intra et inter-classes. Nous étudions les problèmes d'optimisation associés et observons l'effet de certaines variantes de l'algorithme. Dans le second chapitre, nous étudions plus en détail le processus d'apprentissage. Nous présentons le détail des algorithmes intra et inter proposés, ainsi que la façon de les coupler pour l'apprentissage d'un appariement discriminant. Nous mettons en œuvre les algorithmes d'apprentissage présentés sur des données simulées et évaluons leur complexité calculatoire.

Chapitre 3

Apprentissage d'appariements discriminants : formalisation

Les métriques fondées sur la notion d'appariement sont fondamentales pour l'exploration et la discrimination d'ensembles de séries temporelles. Nous proposons, dans ce chapitre, d'apprendre les appariements qui sont les plus discriminants. Nous introduisons un problème d'optimisation fondé sur la maximisation de la variance intra et la minimisation de la variance inter. Nous proposons ensuite une méthode algorithmique pour l'apprentissage des appariements, dont nous déclinons plusieurs variantes.

Comme nous l'avons vu à travers le formalisme rappelé dans le chapitre 1, un grand nombre de travaux visant à comparer deux séries temporelles repose sur la définition d'une distance, ou plus généralement d'une mesure de dissimilarité définie sous la forme d'une fonction de coût (fondée sur les distances entre observations des deux séries) et d'un sous-ensemble d'observation (en général un alignement des arêtes entre les deux séries). Ces mesures de dissimilarité sont utilisées notamment pour des tâches de classification, en général de type "k plus proche voisins" où la série est affectée à la classe à laquelle appartiennent les k séries qui lui sont les plus proches.
Dans la plupart des travaux actuels, la notion d'alignement est au cœur de ces métriques. Dans certains cas, l'alignement est imposé. C'est le cas notamment du couplage euclidien, reliant les observations correspondant aux mêmes instants, et le couplage complet, reliant entre elles toutes paires d'observations des deux séries. Dans d'autres cas, chaque paire de série est alignée en réponse à un problème d'optimisation. Par exemple, dans le contexte de la DTW, le critère de sélection de l'alignement optimal est la minimisation d'une fonction de coût.
Ces alignements usuels font des présuppositions quant aux liens entre les instants. Par exemple, le couplage "euclidien", qui est un des couplages les plus utilisés, part de l'hypothèse que chaque instance est indépendante du passé et du futur, et que les événements correspondants se produisent aux mêmes dates. Le couplage "complet" fait l'hypothèse que tous les couples jouent le même rôle et ignore toute notion de date (pas de structure temporelle). Le couplage induit par la DTW fait l'hypothèse d'une déformation non linéaire du temps, mais respectant les conditions de monotonie, d'exhaustivité et d'extrémités énoncées au chapitre 1

de la partie I. Cependant, ces alignements ne sont pas appris sur la base de caractéristiques communes à la classe. La notion d'alignement est souvent intrinsèque aux paires de série.
Comme nous l'avons soulevé dans les chapitres précédents, notre objectif est de proposer un appariement des séries sur la base de caractéristiques discriminantes au sein de la classe, et non pas un appariement paire à paire. En particulier, certains critères classiques en discrimination sont fondés sur les notions de variances intra-classe et inter-classes. Nous proposons d'apprendre des appariements comme solutions d'un problème d'optimisation inspiré par ces variances, à travers les généralisations proposées au chapitre précédent.
Ce chapitre vise à définir un problème d'optimisation pour la recherche de structures discriminantes au sein d'un ensemble de séries temporelles. Les appariements appris le sont à travers un algorithme d'apprentissage que nous proposons. Nous présentons cet algorithme et un ensemble de variantes, répondant à diverses contraintes ajoutées au problème d'optimisation. Dans un contexte de discrimination, nous cherchons donc à apprendre des appariements, afin de minimiser la variance intra-classe et maximiser la variance inter-classes sous certaines contraintes de normalisation.

1 Introduction au problème d'optimisation lié à la variance d'un ensemble de séries temporelles

Notre problème vise à chercher un ensemble d'appariements entre toutes les paires de séries, afin de minimiser la variance intra et de maximiser la variance inter. Nous allons, dans cette section, mettre en lumière un ensemble de contraintes recherchées pour une définition intéressante des appariements.

1.1 Minimisation de la variance intra

Considérons dans un premier temps le problème de minimisation de la variance intra. Rappelons la formule générale de la variance intra V_W.

$$V_W = \frac{1}{nT} \sum_{k=1}^{K} \sum_{l \in C_k} \sum_{i=1}^{T} \left(x_i^l - \sum_{l' \in C_k} \sum_{i'=1}^{T} \frac{W_{ii'}^{ll'}}{\sum_{\bar{l} \in C_k} \sum_{\bar{i}=1}^{T} W_{\bar{i}\bar{i}}^{l\bar{l}}} x_{i'}^{l'} \right)^2$$

où

$$W^{ll'} = \begin{cases} \lambda \mathbf{I} & \text{si } l = l' \\ \neq \mathbf{0} & \text{si } y_l = y_{l'} \text{ et } l \neq l' \\ \mathbf{0} & \text{si } y_l \neq y_{l'} \end{cases} \tag{25}$$

L'ensemble $\left\{ \frac{W_{ii'}^{ll'}}{\sum_{\bar{l} \in C_k} \sum_{\bar{i}=1}^{T} W_{\bar{i}\bar{i}}^{l\bar{l}}} \right\}$ forme un système de poids ; la variance intra est donc exprimée en fonction des valeurs de ces poids. Nous cherchons une structure d'appariement qui minimise cette variance intra. Certaines conditions, rendues nécessaires pour la définition de la variance telle qu'elle est rappelée ci-dessus, nous permettent de formaliser le problème d'optimisation sous contraintes, que nous représentons de la manière suivante :

$$\begin{cases} \text{Minimiser } V_W \text{ sous les conditions :} \\ \quad \forall k \in \{1,\dots,K\},\ \forall (l,l') \in C_k,\ \forall (i,i') : \\ \quad\quad (i)\ w_{ii}^{ll} > 0 \text{ et } w_{ii'}^{ll} = 0 \text{ pour } i \neq i' \\ \quad\quad (ii)\ \exists (k_1,k_2) w_{k_1k_2}^{ll'} > 0 \end{cases} \tag{26}$$

La condition (i) $w_{ii}^{ll} > 0$ et $w_{ii'}^{ll} = 0$ assure un lien diagonal entre une série et elle-même ; une série est naturellement couplée à elle-même selon un couplage euclidien. Ce couplage est naturel car il s'agit du couplage minimisant la variance entre une série et elle-même. La condition (ii) $\exists (k_1,k_2) w_{k_1k_2}^{ll'} > 0$ assure des blocs non nuls entre séries de même classe ; ceci impose d'avoir au moins un lien entre les séries de chaque couple de séries.
Nous verrons à la section 1.3 que ces deux conditions sont fondamentales pour définir la structure d'appariement, au cœur de la généralisation de la variance. De plus, elles répondent à l'objectif de coupler chaque paire de série, en tenant compte de l'information globale de classe.
Les blocs entre séries de classes différentes sont nuls, puisqu'il n'existe pas de liens structurels entre les classes dans un contexte intra.
Cependant, le problème ci-dessus est mal posé, et les contraintes proposées sont insuffisantes. En effet, l'exemple suivant donne une famille de matrices d'appariement donnant une solution minimale quelle que soit les valeurs prises par les données.

Remarque 35 : (Exemple de structure particulière minimisant le problème)

En notant δ_i^j*, le symbole de Kronecker, qui vaut 1 si* $i = j$ *et* 0 *sinon, on définit la matrice* W
$\forall (l,l') \in C_k, l \neq l'\ w_{ij}^{ll'} = \varepsilon \delta_i^1 \delta_j^1$
Et ainsi,

$$nTV_W = \sum_{\substack{k\in\{1..K\} \\ l\in C_k \\ i\in\{1..T\}}} ((1-n_k\varepsilon)(x_i^l - x_i^l) + \delta_i^1 \sum_{l'\in C_k} \varepsilon(x_1^l - x_1^{l'}))^2$$

Ce qui donne, après réécriture :

$$nTV_W = \sum_{\substack{k\in\{1..K\} \\ l\in\{1..n_k\}}} \left(\sum_{l'=1}^{n_k} \varepsilon(x_1^l - x_1^{l'}) \right)^2$$

En faisant tendre ε vers 0, V_W converge vers 0.

La remarque précédente présente l'exemple d'une matrice d'appariement particulière pour laquelle la variance est aussi faible que souhaitée, quelles que soient les valeurs prises par la série. A partir de conditions limites (la structure diagonale) et en ajoutant du bruit pour se

placer dans les conditions précédentes, nous observons que la variance peut prendre des valeurs aussi faibles que souhaitées. Cet exemple illustre une faiblesse du problème d'optimisation intra-classe. La solution optimale est vide d'information. Il est donc nécessaire de contraindre le problème, pour éviter de converger vers ce cas limite. En particulier, la contrainte (ii) ne suffit pas à imposer un lien réel entre deux séries dans ce cadre.

Les problèmes de minimisation de la variance intra et de maximisation de la variance inter forment deux problèmes qui sont symétriques. Nous présentons à présent le second.

1.2 Maximisation de la variance inter

Nous pouvons, de manière analogue, formuler le problème de maximisation de la variance inter. Rappelons la formule générale de la variance inter V_B.

$$V_B = \frac{1}{nT} \sum_{\substack{k=1\dots K \\ l=1\dots n_k \\ i=1\dots T}} \left(x_i^l - \left(\frac{B_{ii}^{ll}}{\sum_{\substack{\bar{l}=1^n \\ \bar{i}=1^T}} B_{i\bar{i}}^{l\bar{l}}} x_i^l + \sum_{\substack{k' \neq k \\ l'=1^{n_{k'}} \\ i'=1^T}} \frac{B_{ii'}^{ll'}}{\sum_{\substack{\bar{l}=1^n \\ \bar{i}=1^T}} B_{i\bar{i}}^{l\bar{l}}} x_{i'}^{l'} \right) \right)^2$$

avec

$$M^{ll'} = \begin{cases} \lambda \mathbf{I} & \text{si } l = l' \\ \mathbf{0} & \text{si } y_l = y_{l'} \text{ et } l \neq l' \\ \neq \mathbf{0} & \text{si } y_l \neq y_{l'} \end{cases} \tag{27}$$

La variance inter est exprimée en fonction des valeurs des poids $B_{ii'}^{ll'}$. Nous cherchons la structure d'appariement qui maximise cette variance. Nous formalisons à nouveau le problème d'optimisation, que nous considérons comme un problème d'optimisation sous contraintes, représenté de la manière suivante :

$$\begin{cases} \text{Maximiser } V_B \text{ sous les conditions :} \\ \quad \forall k \in \{1, \dots, K\},\ \forall l \in C_k, \forall l' \notin C_k,\ \forall (i, i') : \\ \qquad (i)\ b_{ii}^{ll} > 0 \text{ et } b_{ii'}^{ll} = 0 \text{ pour } i \neq i' \\ \qquad (ii)\ \exists (k_1, k_2) w_{k_1 k_2}^{ll'} > 0 \end{cases} \tag{28}$$

De la même façon, en l'état, ce problème admet une solution évidente, à l'image du cas intra. Nous construisons cette solution.

Remarque 36 : (Exemple de structure particulière maximisant le problème)

$$Notons\ \Delta^i_{\max} = \max_{\substack{(i',l')\\ l\neq l'}} |x^l_i - x^{l'}_{i'}| = |x^l_i - x^{l'_{\max(i,l)}}_{i'_{\max(i,l)}}|$$

avec $(i'_{\max(i,l)}, l'_{\max(i,l)})$ *les valeurs maximisant l'écart pour l'indice* (i,l).

$$\forall B\ V_B \leq \sum_{k=1}^{K} \sum_{l\in C_k} \sum_{i=1}^{T} {\Delta^i_{max}}^2$$

Nous pouvons ainsi définir des appariements pour lesquels la variance inter tend vers cette valeur.

A nouveau, il est nécessaire de contraindre le problème pour rechercher des appariements intéressants, mettant en lumière les liens existant entre les séries.

1.3 Structure d'appariement

Au vu des deux exemples énoncés dans la section 1, il est essentiel d'ajouter des contraintes supplémentaires au problème d'optimisation. Cette section consiste à présenter sommairement la nature des matrices d'appariement que nous considérons, ainsi que certaines variantes. Nous présenterons les contraintes supplémentaires ajoutées au problème d'optimisation pour aboutir à ces matrices.

1.3.a Notations

Nous distinguerons dans la suite deux types d'arêtes, les arêtes liant les instants i et i' de deux séries S_l et $S_{l'}$, et affectées d'un poids $M^{ll'}_{ii'}$, et les arêtes liant globalement les instants i de la série S_l aux instants i' pour tous les couples croisant S_l avec une série d'un ensemble Ω_l. Dans la suite, le terme "arête" fera uniquement référence aux premières. Nous les qualifierons d'"arêtes de couple" lorsqu'il y a une ambiguïté, tandis que nous préciserons toujours "arêtes sémantiques" lorsque nous faisons allusion aux secondes.

Nous utiliserons les notations W et B pour décrire respectivement les matrices d'appariement intra et inter qui seront toujours de la forme suivante.

$$\mathbb{W}_{ll'} = \begin{cases} \lambda I_T & \text{si } l = l' \\ \neq 0 & \text{si } y_l = y_{l'} \\ 0 & \text{si } y_l \neq y_{l'} \end{cases} \tag{29}$$

$$\mathbb{B}_{ll'} = \begin{cases} I_T & \text{if } l = l' \\ 0 & \text{if } y_l = y_{l'} \\ \neq 0 & \text{if } y_l \neq y_{l'} \end{cases} \tag{30}$$

Dans certains contextes (en particulier le cas des arêtes sémantiques), plusieurs blocs sont identiques.

Nous utilisons la notation $\mathbb{W}[M^1, ...M^n]$ (respectivement $\mathbb{B}[M^1, ...M^n]$) lorsque tous les blocs

non nuls (i.e., les blocs intra pour $\mathbb{W}[M^1, ...M^n]$ et les blocs inter pour $\mathbb{B}[M^1, ...M^n]$) associés à une série S^l, hormis les blocs diagonaux, sont égaux à la matrice M^l. Ainsi,

$$\mathbb{W}[M^1, ...M^n]_{ll'} = \begin{cases} I_T & \text{if } l = l' \\ M^l & \text{if } y_l = y_{l'} \\ 0 & \text{if } y_l \neq y_{l'} \end{cases} \quad (31)$$

Nous utilisons la notation $\mathbb{W}[M]$(respectivement $\mathbb{B}[M]$) lorsque tous les blocs non nuls, toutes séries confondues, hormis les termes diagonaux, sont égaux à la matrice M. Ainsi,

$$\mathbb{W}[M]_{ll'} = \begin{cases} I_T & \text{if } l = l' \\ M & \text{if } y_l = y_{l'} \\ 0 & \text{if } y_l \neq y_{l'} \end{cases} \quad (32)$$

Ce dernier cas est un cas particulier du précédent ; en effet, $W[M] = W[\underbrace{M, M, ..., M}_{n \text{ fois}}]$.

Nous introduisons trois structures particulières :

- $\mathbb{W}[0]$ (respectivement $\mathbb{B}[0]$) définit la structure d'appariement intra (respectivement inter) dite Identité. Les blocs diagonaux sont tous égaux à l'Identité, tout les appariements entre séries distinctes sont nuls.
- $\mathbb{W}[\mathrm{I}]$ (respectivement $\mathbb{B}[\mathrm{I}]$) définit la structure d'appariement intra (respectivement inter) dite Euclidienne. Les blocs entre séries de la même classe (respectivement de classes différentes) sont tous égaux à l'identité.
- $\mathbb{W}[\mathrm{J}]$ (respectivement $\mathbb{B}[\mathrm{J}]$) définit la structure d'appariement intra (respectivement inter) dite Unitaire. Les blocs entre séries d'une même classe (respectivement de classes différentes) sont tous égaux au bloc unitaire $J_T = \mathbb{1}_T \, {}^T\mathbb{1}_T$.

1.3.b Arêtes de couple ou arêtes sémantiques

Par l'apprentissage, nous cherchons à obtenir une matrice d'appariement qui soit discriminante. Si les matrices d'appariement classiques tendent à répéter une matrice commune entre chaque paire de séries (exemple : $\mathbb{W}[0]$, $\mathbb{W}[\mathrm{I}]$, $\mathbb{W}[\mathrm{J}]$), nous avons vu que la structure de voisinage peut se définir de façon quelconque, tant qu'est respecté le formalisme générale d'une matrice d'appariement :

$$M_{ll'} = \begin{cases} I_T & \text{if } l = l' \\ \neq 0 & \text{if } l \neq l' \end{cases} \quad (33)$$

Ainsi, au cours de l'apprentissage, les deux approches peuvent être considérées :

- apprendre un appariement intrinsèque à chaque couple de série. Cela revient à obtenir un poids d'arête de couple propre à chaque paire de séries.
- rechercher un appariement commun à toutes les séries. Cela revient à apprendre les pondérations des arêtes sémantiques de chaque classe de séries.

Cette deuxième solution de recherche d'arêtes sémantiques se traduit par des contraintes supplémentaires ; le problème d'optimisation initial se formalise par l'ajout de la condition suivante :

Recherche d'arêtes sémantiques

$$\begin{cases} \text{Minimiser } V_W \text{ sous les conditions :} \\ \quad \forall k \in \{1,\dots,K\},\ \forall (l,l',l'') \in C_k,\ \forall (i,i') : \\ \qquad (i) w_{ii}^{ll} > 0 \text{ et } w_{ii'}^{ll} = 0 \text{ pour } \ i \neq i' \\ \qquad (ii) \text{si } \mathbf{l' \neq l''},\ \mathbf{w_{ii'}^{ll'} = w_{ii'}^{ll''}} \\ \qquad (iii) \exists (k_1,k_2)\ w_{k_1k_2}^{ll'} > 0 \end{cases} \tag{34}$$

Remarque 37 :
Notons que l'ensemble des matrice d'appariement est un espace vectoriel de même dimension que $\mathfrak{M}_T(\mathbb{R})^{n(n-1)}$ *et l'ensemble des matrices d'appariement sémantique est un sous-espace vectoriel du précédent.*
Définissons l'opérateur $\bar{.}^l$

$$\bar{.}^l : \mathbb{M}_{nT}(\mathbb{R}) \to \mathbb{M}_T(\mathbb{R}) \text{tel que } \bar{M}^l = \frac{1}{\#\{l' \neq l \backslash M_{ll'} \neq 0\}} \sum_{l' \neq l} M^{ll'} \tag{35}$$

Nous remarquons que les opérateurs pW et pB, définis de la manière suivante,

$$pWM = \mathbb{W}[\bar{M}^1, \bar{M}^2, ...\bar{M}^n] \tag{36}$$

$$pBM = \mathbb{B}[\bar{M}^1, \bar{M}^2, ...\bar{M}^n] \tag{37}$$

sont des projecteurs de l'ensemble des matrices d'appariement dans l'ensemble des matrices d'appariement sémantique.

Nous avons à présent introduit deux problèmes d'optimisation en vue de la discrimination de séries, consistant à minimiser la variance intra et à maximiser la variance inter sous contraintes. Cependant, une structure de voisinage discriminante est une structure qui minimise la variance intra et maximise la variance inter simultanément, et non indépendamment l'une de l'autre. Nous proposons donc dans la suite un apprentissage simultané des matrices d'appariement.

1.3.c Optimisation simultanée des structures intra et inter-classes

Tels qu'ils sont présentés, les deux problèmes d'optimisation initiaux (équations 26 et 28) recherchent des matrices d'appariement W et B optimisant les deux critères. Les deux matrices d'appariement W et B étant définies à partir de blocs matriciels disjoints, il est aisé de proposer des méthodes d'apprentissage séparées pour les deux types de liens. Cependant, dans l'objectif d'apprendre des appariements discriminants, il peut sembler intéressant de lier les deux approches. Dans le cadre de la recherche d'arêtes sémantiques, une manière de lier les deux approches consiste à optimiser simultanément les deux critères, i.e., rechercher un appariement sémantique commun aux deux structures, optimisant une fonction de V_W et de V_B.

recherche d'arêtes discriminantes dans le cadre d'arêtes sémantiques

$$\left\{\begin{array}{l} \text{Minimiser } \mathbf{f}(\mathbf{V_W}, \mathbf{V_B}) \text{ sous les conditions :} \\ \quad \forall(\mathbf{l}, \mathbf{l}', \mathbf{l}''),\ \forall(i, i') : \\ \qquad (i) m_{ii}^{ll} > 0 \text{ et } m_{ii'}^{ll} = 0 \text{ pour } i \neq i' \\ \qquad (ii) \text{si } \mathbf{l}' \neq \mathbf{l}'',\ \mathbf{m_{ii'}^{ll'}} = \mathbf{m_{ii'}^{ll''}} \\ \qquad (iii) \exists(k_1, k_2)\ w_{k_1 k_2}^{ll'} > 0 \end{array}\right. \tag{38}$$

Répondre à notre objectif de discriminer les séries temporelles (i.e., de minimiser la variance intra et de maximiser la variance inter) revient à chercher un appariement entre paires de série, minimisant un critère fondé sur la variance intra et maximisant un critère fondé sur la variance inter. La solution classique, inspirée de l'analyse factorielle discriminante consiste à prendre comme fonction $\mathbf{f}(\mathbf{V_W}, \mathbf{V_B})$ le quotient V_W/V_B.

Dans le cadre d'un apprentissage séparé des deux structures, il est possible de grouper l'apprentissage différemment. Ceci sera développé dans la suite. Nous présentons à présent différentes variantes intervenant dans la définition des matrices d'appariement. La première variante consiste à opter pour une approche soit booléenne, i.e., le lien entre deux instants est actif ou inactif, soit pondérée, i.e., le lien entre deux instants prend une certaine valeur entre 0 et 1 qui quantifie son intensité.

1.3.d Granularité de la structure d'appariement : booléenne ou pondérée

Un bloc matriciel entre un couple de séries, apparaissant dans les matrices d'appariement décrites précédemment, consiste à décrire des liens existant entre les instants de la première série, et ceux de la seconde. Les contraintes décrivant la structure d'appariement imposent d'avoir des blocs non nuls entre les séries liées (séries de la même classe dans le cadre de la structure d'appariement intra, séries de classes différentes dans le cadre de la structure d'appariement inter). Deux types de liens peuvent être considérés.

1. Choix de liens booléens.
   Le lien entre deux arêtes est booléen : deux arêtes sont liées ou non. Cette notion est une généralisation de la notion d'alignement. Au lieu d'imposer le choix d'arêtes contigües, selon les contraintes de monotonie, d'exhaustivité, et d'extrémité, les arêtes sont sélectionnées selon de nouvelles contraintes d'optimalité.

$$\left\{\begin{array}{l} \text{Minimiser } V_W \text{ sous les conditions :} \\ \quad \forall k \in \{1, \ldots, K\},\ \forall(l, l') \in C_k,\ \forall(i, i') : \\ \qquad (i) w_{ii}^{ll} > 0 \text{ et } w_{ii'}^{ll} = 0 \text{ pour } i \neq i' \\ \qquad (ii) \exists(k_1, k_2) w_{k_1 k_2}^{ll'} > 0 \\ \qquad (iii) \text{si } \mathbf{w_{ii'}^{ll'}} \neq \mathbf{0}, \mathbf{m_{ii'}^{ll'}} = \mathbf{m_{ii}^{ll}} \end{array}\right. \tag{39}$$

   Ce choix contraint toutes les arêtes liées à un instant donné à être équi-pondérées. En matière de problème d'optimisation, le fait d'imposer des liens booléens conduit à un problème d'optimisation discret.

2. Choix de liens pondérés.
   Le lien entre deux arêtes est progressif, deux arêtes sont toujours liées avec une intensité plus ou moins forte. Cette notion est une généralisation de la notion de couplage complet. Au lieu d'imposer un poids uniforme sur toutes les arêtes, les arêtes sont pondérées selon de nouvelles contraintes d'optimalité. Ce choix permet de moduler l'intensité des différentes arêtes. En matière de problème d'optimisation, le fait d'autoriser des liens pondérés conduit à un problème d'optimisation convexe.

$$\begin{cases} \text{Minimiser } V_W \text{ sous les conditions :} \\ \quad \forall k \in \{1,\dots,K\},\ \forall (l,l') \in C_k,\ \forall (i,i') : \\ \qquad (i) w_{ii}^{ll} > 0 \text{ et } w_{ii'}^{ll} = 0 \text{ pour } i \neq i' \\ \qquad (ii) \forall (k_1,k_2) w_{k_1k_2}^{ll'} > 0 \end{cases} \tag{40}$$

Si l'approche booléenne est un cas particulier de l'approche progressive, la différence sémantique entre les deux est fondamentale. Dans un cas, nous obtenons un sous-ensemble de liens, c'est une généralisation de la notion d'alignement ; dans le second, nous obtenons un système de poids, c'est une généralisation de la notion de couplage complet. Nous avons présenté deux méthodes d'optimisation, expliquant comment ces deux approches pouvaient être couplées pour la discrimination d'une partition de séries temporelles. Dans la suite, dans le cadre de ces deux problèmes d'optimisation, nous introduisons une méthode algorithmique permettant d'apprendre des appariements qui vérifient ces contraintes et visent à s'approcher d'une solution aux deux problèmes.

2 Présentation d'une méthode pour l'apprentissage des appariements.

Nous présentons dans cette section une méthode d'apprentissage d'appariements pour résoudre les deux problèmes d'optimisation introduits dans la section précédente.
Nous définissons en amont un critère de sélection d'arêtes, en lien avec notre problème d'optimisation et nous initialisons le processus avec une matrice donnée.
Cette méthode est fondée sur un processus itératif. A chaque itération, nous sélectionnons une arête dont le poids est modifié, jusqu'à stabilité.
Notre approche vise à pénaliser certaines arêtes et à en renforcer d'autres ; nous ajoutons une étape de renormalisation de la matrice qui permet une redistribution des modifications.

2.1 Proposition d'une méthode d'apprentissage des structures de voisinage

Algorithme

La structure générale de la méthode est la suivante. Les figures 15 et 16 illustrent cette approche, en réponse aux deux types de problèmes d'optimisation, soit booléen, soit pondéré.

Les deux méthodes reposent sur le même schéma décrit dans l'algorithme 1. La différence entre les deux approches repose dans l'intensité des pénalisations. La structure progressive conduit à des appariements ayant moins d'arêtes couplées. Nous pouvons remarquer que la

Algorithm 1 Structure générale de la méthode d'apprentissage

1: Initialisation : définition de la matrice initiale
2: **repeat**
3: Définition du critère de sélection des arêtes
4: Mise à jour de la matrice d'appariement
5: **until** critère de fin
6: **return** La matrice d'appariement obtenue

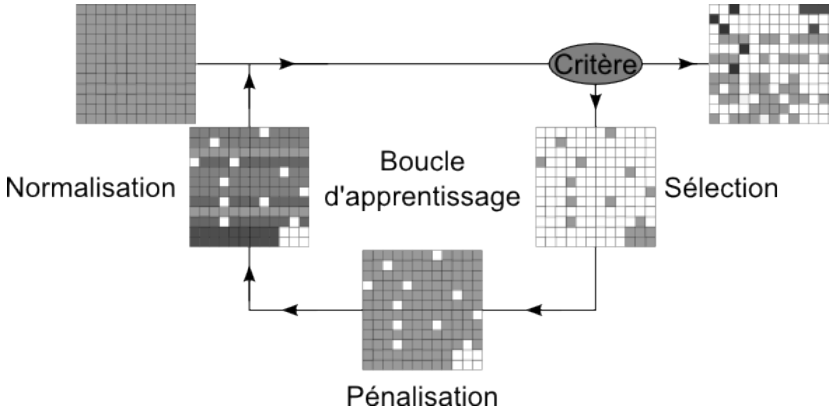


FIGURE 15 – Schéma récapitulatif de la méthode booléenne

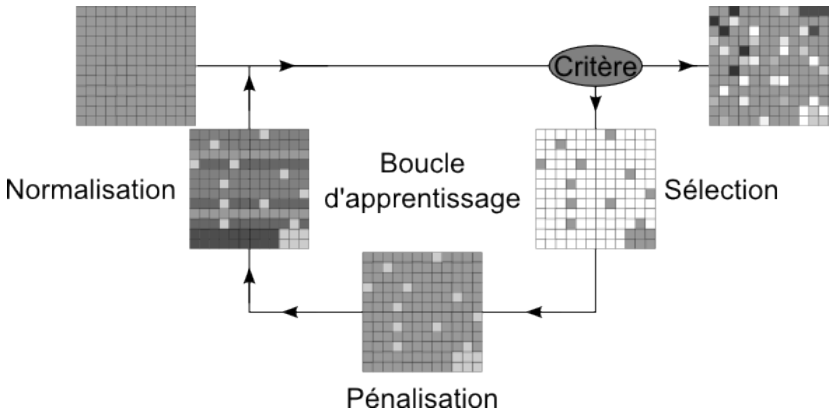


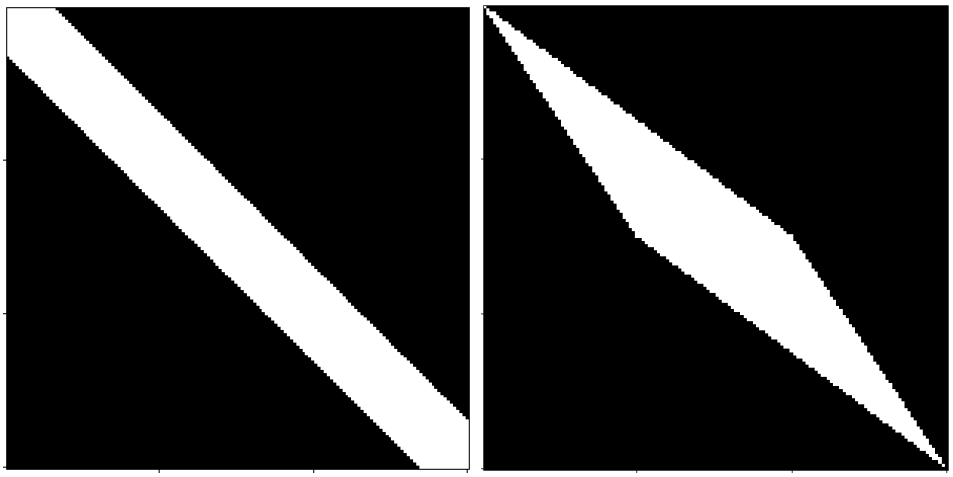
FIGURE 16 – Schéma récapitulatif de la méthode pondéré

structure booléenne présente des couples déconnectés, tandis que la structure progressive conserve toutes les arêtes. Dans la suite, nous présentons pour chacune des étapes de l'algorithme ci-dessus, les différentes approches envisagées.

La première étape qui intervient dans la méthode proposée est l'initialisation de l'algorithme. Nous montrons que cette étape est ici fondamentale dans le processus d'apprentissage.

2.2 Définition de la matrice initiale $\mathbb{W}$

Le choix de la matrice initiale est fondamental, car il conditionne l'apprentissage de la structure de voisinage, en déterminant les zones accessibles. La matrice WInitiale décrit le poids des voisins à l'initialisation du processus d'apprentissage. Lorsqu'on a au départ une information a priori sur ces poids, on peut privilégier certaines arêtes par rapport à d'autres. La structure d'appariement associée à cette hypothèse est celle qui lie les séries selon un couplage complet, compatible avec la structure d'appariement initiale, i.e., dans le cas d'une matrice d'appariement intra-classe, la structure $\mathbb{W}[J]$, et dans le cas inter-classes, la structure $\mathbb{B}[J]$. Cette matrice initiale peut être également fixée arbitrairement, de sorte que les appariements entre séries se limitent à des instants voisins. Cette initialisation n'est pas sans rappeler les contraintes globales proposées pour la DTW. Dès lors que les blocs sont fidèles aux structures d'appariements définies ci-dessus dans les formules 29 ou 30, nous pouvons définir chaque bloc initial de façon arbitraire.

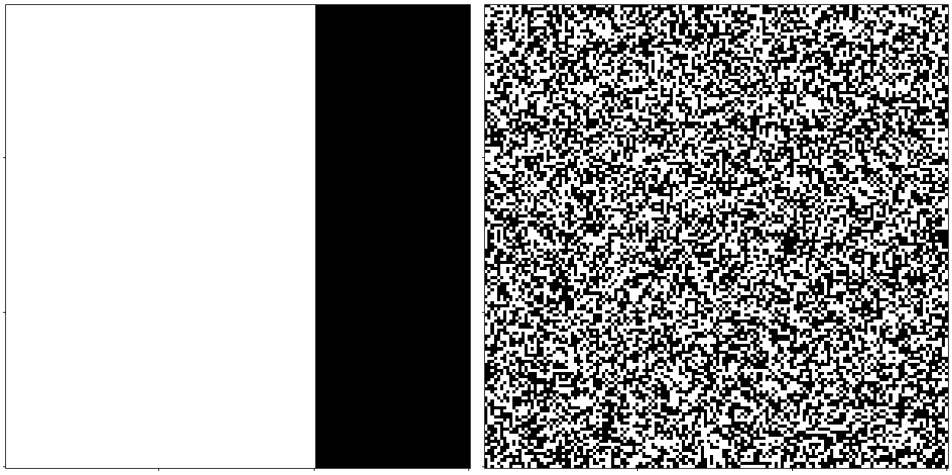


(a) Bande de Sakoe-Chiba (b) Parallélogramme d'Itakura

FIGURE 17 – Appariements particuliers avec écarts fixés autour de la diagonale

La figure 17 présente plusieurs exemples d'appariements initiaux centrés autour de la diagonale. Ce type de contraintes est fréquemment utilisé dans le cadre des alignements, pour éviter de relier entre eux des instants qui ne se correspondent pas (Sakoe et Chiba, 1978; Itakura, 1975). Ces structures de voisinage particulières sont également compatibles avec la notion d'appariement.

D'autres types d'appariements initiaux peuvent être proposés. Nous pouvons par exemple considérer, dans le cadre d'une prédiction précoce, de ne nous intéresser qu'aux liens avec des instants antérieurs à une date fixée 2.2(a). Dans d'autres cas, nous pouvons initialiser l'apprentissage à partir d'une matrice initiale préalablement calculée, par exemple solution



(a) Prédiction précoce (b) Structure initiale donnée

FIGURE 18 – Appariements répondant à des problèmes particuliers

d'un problème d'optimisation annexe 2.2(b).

Le choix de la matrice initiale offre un choix de modifications très large. Les arêtes étant choisies au départ, l'ensemble des arêtes sélectionnées en vue de la pénalisation peut varier. Dans le chapitre 4 de cette partie, nous utilisons la matrice d'initialisation pour coupler les appariements intra et inter-classes. Celle-ci joue un rôle essentiel dans l'approche discriminante que nous proposons.

A l'issue du choix de la matrice initiale, un procédé itératif se met en place. Nous présentons à présent la manière d'introduire le critère d'optimisation au sein de l'algorithme, pour l'évolution du problème de discrimination. Plusieurs approches sont envisagées et discutées afin de choisir les arêtes qui seront pénalisées ou renforcées.

2.3 Définition du critère à optimiser

La méthode proposée pour l'apprentissage consiste à sélectionner au départ un ensemble d'arêtes d'intérêt, puis de choisir dans cet ensemble celles qui sont à renforcer ou à pénaliser. Nous proposons plusieurs quantités fondées sur la variance, et plus particulièrement la variation de la variance lors de la modification des poids d'une arête, pour la sélection d'arêtes.

Dérivée selon $m_{ii'}$ La dérivée de la variance-covariance en fonction du poids des arêtes consiste à effectuer de petites variations dans chacune des directions des poids de la matrice ; l'objectif est de calculer l'impact de cette variation. Nous pouvons à cet effet calculer la dérivée partielle de la variance de Mom selon $m_{ii'}$.

$$\frac{\partial V_M}{\partial m_{ii'}^{ll'}} = \frac{-2x_{i'}^{l'}}{n}\left(x_i^l - \sum_{i_1,l_1} m_{ii_1}^{ll_1} x_{i_1}^{l_1}\right)$$

Le fait de chercher à optimiser la structure de voisinage par cette méthode s'apparente à la méthode du gradient projeté. Cette méthode présente l'inconvénient de faire le calcul de la variance en amont des contraintes. En effet, le choix de la direction optimale se fait avant

la normalisation. Cette dernière entraîne ensuite une modification de l'espace des données, et n'assure pas l'optimalité. Ainsi, ce calcul ne tient pas compte de la spécificité de l'expression de la variance, définie sous des contraintes de normalisation en ligne de chaque observation.

Pour tenir compte de ces contraintes dans la sélection des arêtes, nous proposons donc dans ce qui suit deux approches tenant compte de l'impact ressenti lors de la suppression d'une arête, et prenant en considération la normalisation. Dans notre objectif de minimiser la variance intra, nous définissons, pour chaque arête, les notions de contribution à la variance et de potentiel discriminant, qui est la contribution au pouvoir discriminant d'une arête. Ces deux notions s'expriment comme un différentiel des valeurs prises par une grandeur au moment de la suppression d'une arête.

2.3.a Contribution d'une arête à la variance.

Soit $\mathbb{M}$ une matrice d'appariement. On appelle contribution d'une arête à la variance le différentiel entre la variance associée au voisinage $\mathbb{M}$ et la variance associée au voisinage $\mathbb{M}\|ii'$ lorsque l'arête est supprimée. En effet, la définition de la variance fondée sur une structure d'appariement nécessite la normalisation en ligne de la matrice d'appariement. Ainsi, supprimer une arête répartit l'intensité de son poids sur les autres arêtes.

Définition 38 : (Contribution)
La contribution de l'arête i,l,i',l' notée $CM_{ii'}^{ll'}$ *vaut* $Var(\mathbb{M}) - Var(\mathbb{M}|i,l,i',l')$.
La contribution de l'arête sémantique i,i' notée $CM_{ii'}$ vaut $Var(\mathbb{M}) - Var(\mathbb{M}|i,i')$.

Le schéma suivant explique les étapes implicites lors du calcul de la variance associée à la suppression d'une arête.

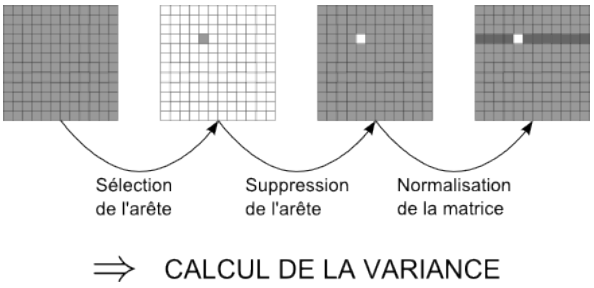


FIGURE 19 – Calcul de la variance tronquée

Remarque 39 : (Existence de la contribution)
L'existence de la contribution est uniquement liée à la possibilité de calculer la variance tronquée $Var(\mathbb{M}|i,l,i',l')$*, correspondant au terme de variance après suppression de l'arête* (i,l,i',l')*. Cette existence découle donc uniquement de la possibilité de normaliser la matrice d'appariement en ligne. Celle-ci ne pouvant se faire que si la condition* $\sum_{i_1,l_1} m_{ii_1}^{ll_1} \neq 0$ *est vérifiée, cela revient, du fait de la positivité des poids au cœur de la structure d'appariement, à s'assurer de l'existence d'un élément* $m_{ii_0}^{ll_0}$ *non nul pour chaque instant (i,l). Ce point est assuré par la nature même des matrices d'appariement qui assignent un poids non nul à la diagonale de la matrice.*

1. Une arête dont la contribution est positive est une arête dont le différentiel de variance à l'issue de sa suppression est positif.
   La suppression de cette arête entraîne une diminution de la variance. C'est donc une arête qui augmente la variabilité de la structure.
2. Une arête dont la contribution est négative est une arête dont le différentiel de variance à l'issue de sa suppression est négatif.
   Sa suppression entraîne une augmentation de la variance. C'est donc une arête qui réduit la variabilité de la structure.

Plus formellement,

Propriété 40 :

$$CM_{ij} > 0 \iff \underbrace{Var(\mathbb{M})}_{\text{variance avec l'arête}} > \underbrace{Var(\mathbb{M}|ij)}_{\text{variance sans l'arête}}$$

La contribution d'une arête caractérise sa tendance à apporter de la variabilité à la structure.

Proposition 41 : (Calcul de la contribution par la formule des centres mobiles)

$$CM_{ii'}^{ll'} = \sum_{j=1}^{p} \underbrace{\frac{-M_{ii'}^{ll'}}{1-M_{ii'}^{ll'}}(MX_{ij}^{l} + X_{i'j}^{l'})}_{\text{moyenne de } X \text{ et de } MX} \underbrace{(2X_{ij}^{l} - \frac{2-M_{ii'}^{ll'}}{1-M_{ii'}^{ll'}} MX_{ij}^{l})}_{\text{écart entre } X \text{ et } MX}$$

En particulier, plus une arête a un poids fort, plus elle contribue à la variance. Au contraire, une arête ayant un poids nul a une contribution nulle. La réciproque est fausse.

Remarque 42 :

$$M_{ii'}^{ll'} = 0 \Longrightarrow CM_{ii'}^{ll'} = 0$$

Ensemble des arêtes apportant de la variabilité Dans le cadre d'un problème de discrimination, les arêtes qu'on cherche à conserver sont celles qui relient des zones homogènes au sein de la structure intra-classe, tandis qu'on souhaite pénaliser les arêtes qui relient des zones homogènes au sein de la structure inter-classes. Nous définissons l'ensemble $\mathfrak{A}^{+}[M]$ des arêtes d'intérêt, comme l'ensemble des arêtes dont la contribution est négative lorsqu'on cherche à minimiser la variance, positive lorsqu'on cherche à la maximiser.

$$\mathfrak{A}^+[W] = \{(i,l,i',l') \in \mathfrak{A} / CW_{ii'}^{ll'} < 0\}$$
$$\mathfrak{A}^+[B] = \{(i,l,i',l') \in \mathfrak{A} / CB_{ii'}^{ll'} > 0\}$$

On définit l'ensemble "pioche" $\mathfrak{A}^-$, comme l'ensemble des arêtes à pénaliser. En général, l'ensemble pioche est inclus dans le complémentaire de l'ensemble d'intérêt au sein des arêtes de poids non nuls, c'est-à-dire l'ensemble des arêtes dont la contribution est positive lorsqu'on veut minimiser la variance, et négative lorsqu'on cherche à la maximiser.

$$\mathfrak{A}^-[W] \subseteq \{(i,l,i',l') \in \mathfrak{A} / CW_{ii'}^{ll'} > 0\}$$
$$\mathfrak{A}^-[B] \subseteq \{(i,l,i',l') \in \mathfrak{A} / CB_{ii'}^{ll'} < 0\}$$

2.3.b Potentiel discriminant d'une arête sémantique.

Dans le cadre d'un apprentissage de poids d'arêtes sémantique, nous définissons le potentiel discriminant d'une arête sémantique. Rappellons tout d'abord la notion de pouvoir discriminant.

Définition 43 : (Pouvoir discriminant)
Le pouvoir discriminant associé à une structure de voisinage et noté ρ vaut $\frac{Var(\mathbb{W})}{Var(\mathbb{W})+Var(\mathbb{B})}$

On appelle potentiel discriminant d'une arête sémantique ii' l'évolution de son pouvoir discriminant ρ lors de la suppression de l'arête sémantique.

Définition 44 : (Potentiel discriminant)
Le potentiel discriminant de l'arête sémantique ii' notée $C\rho_{ii'}$ vaut $\frac{Var(\mathbb{W})}{Var(\mathbb{W})+Var(\mathbb{B})} - \frac{Var(\mathbb{W}|ii')}{Var(\mathbb{W}|ii')+Var(\mathbb{B}|ii')}$.

Remarquons que dans le cas statique, la quantité $Var(\mathbb{W})+Var(\mathbb{B})$ est égale à la variance totale. Ici, du fait d'une structure en graphe non complet, le résultat n'est plus assuré.

Remarque 45 : (Existence du potentiel discriminant)
A l'instar de la contribution d'une arête, l'existence du potentiel discriminant découle de la possibilité de calculer la variance tronquée $Var(\mathbb{M}|i,l,i',l')$; ce point est assuré par la nature des matrices d'appariement.

Le potentiel discriminant est donc défini comme un différentiel entre les quotients des variances intra et inter, avant et après suppression de l'arête sémantique.

Remarque 46 : (Réécriture du pouvoir discriminant)
Notons que le pouvoir discriminant ρ est lié au quotient des variances intra et inter par la formule suivante :

$$\rho = \frac{Var(\mathbb{W})}{Var(\mathbb{W}) + Var(\mathbb{B})} = \frac{1}{1 + \frac{Var(\mathbb{B})}{Var(\mathbb{W})}}$$

Nous obtenons alors la proposition suivante :

Proposition 47 :

$$C\rho_{ii'} > 0 \iff \underbrace{\frac{Var(\mathbb{W})}{Var(\mathbb{B})}}_{\rho_{ii'}\ avec\ l'arête} > \underbrace{\frac{Var(\mathbb{W}|ii')}{Var(\mathbb{B}|ii')}}_{\rho_{ii'}\ sans\ l'arête}$$

Comme son nom l'indique, le potentiel discriminant d'une arête sémantique définit sa tendance à discriminer les séries.

1. Une arête sémantique dont le potentiel discriminant est positif est une arête dont la suppression entraîne une augmentation du rapport variance intra/inter.
2. Une arête sémantique dont le potentiel discriminant est négatif est une arête dont la suppression entraîne une diminution du rapport variance intra/inter.

Ceci donne, en particulier, les implications suivantes

$$C\rho_{ii'} > 0 \Longrightarrow CW_{ii'} > 0 \text{ ou } CB_{ii'} < 0$$
$$C\rho_{ii'} < 0 \Longrightarrow CW_{ii'} < 0 \text{ ou } CB_{ii'} > 0$$

Ensemble des arêtes apportant de la variabilité Nous définissons l'ensemble $\mathfrak{A}^+[\rho]$ des arêtes sémantiques d'intérêt, comme l'ensemble des arêtes sémantiques dont le potentiel discriminant est négatif.

$$\mathfrak{A}^+[\rho] = \{(i,l,i',l') \in \mathfrak{A} / C\rho_{ii'} < 0\}$$

On définit l'ensemble "pioche" $\mathfrak{A}^-[\rho]$, comme l'ensemble des arêtes sémantiques à pénaliser. Souvent, on considère comme ensemble pioche le complémentaire de l'ensemble d'intérêt au sein des arêtes de poids non nuls, c'est à dire, l'ensemble des arêtes sémantiques dont le potentiel discriminant est strictement positif.

$$\mathfrak{A}^-[\rho] \subseteq \{(i,l,i',l') \in \mathfrak{A} / C\rho_{ii'} > 0\}$$

les méthodes proposées ci-dessus ont toutes leurs spécificitées. Nous allons voir les modifications qu'elles induisent quant au problème d'optimisation initial.

2.3.c Incidence sur le problème d'optimisation

Outre le fait de modifier le critère à optimiser, le choix de l'approche "contribution" ou "potentiel discriminant" impacte le problème d'optimisation considéré. Il s'avère notamment nécessaire de choisir l'approche fondée sur les arêtes sémantiques pour considérer le potentiel discriminant (cf discussion section 1.3.c). Notons que, dans le cas du potentiel discriminant,

nous optimisons une fonction particulière des variances intra et inter. Dans le cas de la contribution, les problèmes d'optimisation sont séparés, et le lien existant entre l'intra-classe et l'inter-classes doit être introduit en sus.

Après la présentation de toutes ces variantes, nous allons maintenant présenter et justifier le choix qui a été fait dans la définition d'une méthode d'apprentissage dans le cadre de cette thèse.

2.4 Choix de l'approche

Nous avons présenté plusieurs variantes de cet algorithme. Nous allons dans cette partie discuter de toutes les spécificités de ces variantes. Nous allons, dans un premier temps, discuter de la différence entre arêtes de couples et arêtes sémantiques.

2.4.a Arêtes de couple

Le fait d'apprendre un bloc sémantique permet de définir un algorithme plus rapide, et moins coûteux en ce qui concerne la mémoire. En effet, il évite toutes les mises à jour de matrices individuelles et ne stocke qu'un seul bloc matriciel par classe. De plus il est possible d'intégrer directement à l'algorithme l'apprentissage de blocs différentiels à travers le potentiel discriminant.
En revanche, l'apprentissage est moins précis ; en effet, d'une série à l'autre, la position globale d'un événement peut varier, et considérer les arêtes sémantiques ne permet pas de différencier les séries entre elles. Ce point est contraire à l'idée que des séries peuvent varier au sein d'une même classe.
Nous avons donc choisi, dans le cadre de cette thèse, d'apprendre des blocs qui définissent un couplage entre paires de séries, malgré son aspect plus coûteux en termes de complexité.

De ce fait, nous avons abandonné l'approche fondée sur le potentiel discriminant (évoquée en 2.3.b), qui est basée nécessairement sur l'apprentissage d'arêtes sémantiques.
L'approche choisie consiste à rechercher la structure de voisinage qui optimise la contribution d'une arête à la variance. Le problème d'optimisation associé est rappelé ci-dessous, dans l'équation 41 : il correspond à une modification du problème initial, dans le contexte de la variante progressive. Les modifications inhérentes à l'apprentissage booléen sont ajoutées en gris.

$$\left\{\begin{array}{l} \text{Minimiser } V_W \text{ sous les conditions :} \\ \quad \forall k \in \{1,\dots,K\},\ \forall (l,l') \in C_k,\ \forall (i,i') : \\ \qquad (i) w_{ii}^{ll} > 0 \text{ et } \ w_{ii'}^{ll} = 0 \text{ pour } i \neq i' \\ \qquad (ii) \exists (k_1,k_2) w_{k_1k_2}^{ll'} > 0 \\ \qquad (iii) \text{si } \mathbf{w_{ii'}^{ll'}} \neq \mathbf{0}, \mathbf{m_{ii'}^{ll'}} = \mathbf{m_{ii}^{ll}} \end{array}\right. \tag{41}$$

Le choix d'arêtes de couples conduit en particulier au choix de la contribution des arêtes à la variance comme critère d'optimisation.

2.4.b Contribution d'une arête à la variance

La plupart des méthodes d'appariement classiques cherchent à coupler les séries deux par deux. Les alignements type DTW, par exemple, recherchent un chemin entre les instants de S^l et de $S^{l'}$ minimisant leur écart. Le recours à la variance permet d'impliquer l'ensemble des séries dans le choix des arêtes, et de pénaliser les arêtes selon un critère global, au contraire d'un critère paire à paire. Nous pouvons alors trouver des éléments communs à la classe au cœur d'un couple de séries très éloignées.

La contribution proportionnelle au poids de l'arête A partir de la formule des centres mobiles donnée à la proposition 41, il apparaît une proportionnalité entre le poids M_{ij} et la contribution CM_{ij}. Ainsi, lorsque deux arêtes ont un poids différent, l'arête au poids le plus important aura une contribution plus forte. Cet aspect conduit l'apprentissage à réduire rapidement le poids des arêtes les plus fortes quand celles-ci contribuent positivement à la variance.

Nous avons choisi de fonder notre approche sur la contribution d'une arête à la variance. Nous voulons voir en quoi ce choix répond à notre problème de discrimination. Nous nous penchons donc de manière théorique sur l'incidence qu'a la pénalisation d'une arête en fonction de sa contribution.

2.5 Impact sur la variance de la pénalisation d'une arête

Nous nous concentrons uniquement sur le cas de la pénalisation. Nous justifierons ce choix dans la suite.
Nous sélectionnons les arêtes d'intérêt en fonction des répercussions qu'auraient leur suppression. L'idée intuitive derrière cette stratégie est la suivante : *si la suppression d'une arête entraîne une diminution ou une augmentation de la variance, l'effet de la pénalisation de cette arête va en général dans le même sens.* Nous présentons en annexe une preuve de ce point. Cette preuve consiste à étudier le signe de la variance en fonction de l'intensité de la pénalisation. La démonstration se limite au cadre de la pénalisation d'une seule arête. Si plusieurs arêtes sont pénalisées simultanément, le résultat n'est plus assuré.

Proposition 48 : (Signe de la contribution)
En notant $\Delta_{ii'}(\beta) = V - V_{ii'}(\beta)$, cette expression est du même signe que l'expression

$$(1-\beta)(\bar{x}^i - x_j - \beta p_{ij} x_j)$$
$$(2(1-(1-\beta)p_{ij})x_i - (2-(1-\beta)p_{ij})\bar{x}^i - (\beta-1)(1+(\beta+2)p_{ij})p_{ij}x_j)$$

qui est un polynôme de degré 4, dont 1 et $\frac{\bar{x}^i - x_j}{p_{ij}x_j}$ sont deux racines évidentes. Nous pouvons donc exprimer les racines et déduire aisément le signe de la fonction.

La démonstration explicite se trouve à l'annexe B 1, l'étude du signe de cette fonction est menée dans la section 2.

2.6 Mise à jour de la matrice d'appariement

La méthode d'apprentissage mise en œuvre consiste à mettre à jour de façon itérative la matrice d'appariement. Les figures 15 et 16 rappellent le cœur de l'approche proposée pour répondre au problème d'optimisation : une première étape de sélection, puis la pénalisation d'arêtes sélectionnées, puis la renormalisation de toutes les arêtes.

- Problème booléen : cela revient à désactiver certaines arêtes.
- Problème pondéré : cela revient à diminuer les poids de certaines arêtes et à renforcer le poids d'autres.

En outre, notons que le calcul de la variance impose une normalisation en ligne des matrices au préalable. Les poids des arêtes sont relatifs. La pénalisation du poids des arêtes de la pioche induit donc un renforcement des arêtes d'intérêt. Par symétrie, le renforcement des arêtes d'intérêt conduit à une pénalisation des autres arêtes.

Les critères de sélection des arêtes sont des critères différentiels. Ils sont liés à l'impact de la suppression d'une arête.

- Approche booléenne : rappelons (cf. remarque 42) que le poids nul affecté à certaines arêtes est irréversible. Le renforcement n'a pas de sens dans ce cadre là.
- Approche pondérée : la pénalisation et le renforcement sont complémentaires. La pénalisation de certaines arêtes entraîne le renforcement des autres lors de la renormalisation. Réciproquement, le renforcement d'une arête entraîne la pénalisation des autres. La différence entre les deux actions repose dans le niveau de pénalisation. Dans le premier cas, nous avons une influence sur les poids des arêtes renforcées, dans le second, une influence sur les poids des arêtes pénalisées. Dans le cadre de la pénalisation, toutes les arêtes renforcées le sont avec un poids équivalent, tandis que dans le cadre du renforcement, ce sont les arêtes pénalisées qui le sont avec un poids équivalent.

Du fait de la complémentarité de ces deux aspects et de leurs nombreuses similarités, nous faisons le choix d'une méthode fondée uniquement sur la pénalisation des arêtes. L'approche que nous considérons est donc une pénalisation du poids des arêtes appartenant à l'ensemble pioche. Nous justifions à présent ce choix.

Pénalisation ou renforcement Le renforcement et la pénalisation sont deux approches complémentaires liées par la normalisation de la matrice. Leur différence repose sur le fait que la pénalisation des arêtes de la pioche induit une modulation des poids des arêtes pénalisées, et un renforcement uniforme des arêtes d'intérêt par renormalisation ; c'est l'opposé dans le cas du renforcement.

Dans le cas de la minimisation de la variabilité intra, le fait pour une arête d'apporter de la variabilité (cas des arêtes à contribution positive) est quantifiable. La pénalisation tend à homogénéiser le voisinage. Plus l'arête augmente la variabilité, plus on souhaite la pénaliser. Au contraire, le fait de renforcer une arête d'intérêt n'est pas judicieux. Une arête dont la contribution est négative est une arête qui n'apporte pas autant de variabilité que d'autres. Il n'y a cependant pas de raisons de la privilégier a priori. Les poids les plus négatifs ne correspondent pas forcément aux arêtes les plus intéressantes, mais à celles dont l'impact sur la décroissance est le plus important, relativement à la structure d'appariement. Dans le cas de la maximisation de la variabilité inter, par symétrie, plus l'arête diminue la variabilité, plus on souhaite la pénaliser. La valeur de la contribution d'une arête d'intérêt ne donne pas

d'information sur l'hétérogénéité qu'elle apporte. De plus, le calcul de la contribution d'une arête est une approche différentielle visant à observer l'impact de la suppression d'une arête. La modification de la variance lors de la pénalisation d'une arête, comme nous le démontrons plus tard dans le chapitre, va en général dans le même sens que lors de la suppression définitive.

Ainsi, la pénalisation est plus naturelle que le renforcement. Nous nous limiterons donc dans notre approche, à une pénalisation des arêtes de la pioche.

Pour la sélection des arêtes à pénaliser au sein de la pioche, nous proposons plusieurs approches, qui ont toutes une sémantique très différente. Les différentes variantes proposées consistent à faire varier l'intensité de la pénalisation, le nombre d'arêtes qui sont pénalisées à chaque étape, et la sévérité de l'approche, à travers la manière dont sont impactées les modifications sur les arêtes à renforcer et les critères d'arrêt du processus itératif.

2.6.a Stratégie de pénalisation

Nous présentons ainsi plusieurs stratégies au cœur du processus de pénalisation. Ces variantes reposent sur la manière de pénaliser les arêtes selon les trois aspects d'intensité, cardinalité et sévérité présentés ci-dessus.

Intensité des pénalisations Si les deux problèmes d'optimisation booléens et progressifs présentent des différences fondamentales en ce qui concerne la signification, l'approche mise en œuvre pour trouver une solution optimale est pourtant très semblable pour les deux. L'intensité des pénalisations constitue la principale différence entre les deux approches.

- cas du problème d'optimisation booléen :
  la modification du poids de l'arête consiste à mettre à 0 le poids de l'arête à pénaliser. L'intensité de la pénalisation est maximale.
- cas du problème d'optimisation pondéré :
  C'est le cas général, le problème booléen en est un cas limite. Nous adoptons une pénalisation progressive. La pénalisation d'une arête repose sur le choix d'un taux de pénalisation. Le choix de l'intensité de la pénalisation a un impact important sur la vitesse de convergence, mais également sur la finesse de l'apprentissage. Plus l'intensité est forte, plus les chances de voir une arête pénalisée à un instant donné être affectée d'un poids important se réduit. Le choix de poids forts rend l'approche plus longue à converger.
  Nous pouvons choisir pour cette approche un taux de pénalisation fixé ou choisir une pénalisation fondée sur la valeur de la contribution, en considérant que plus une arête contribue fortement à augmenter la variance, plus elle doit être pénalisée. De même, nous pouvons choisir de faire de petites modifications à chaque fois, ou des modifications plus fortes.

Approche globale ou locale L'approche globale consiste à pénaliser à chaque itération toutes les arêtes appartenant à la pioche. Au contraire, l'approche locale consiste à cibler à chaque itération une arête au sein de l'ensemble pioche et à la pénaliser. L'arête est sélectionnée de manière aléatoire ou arbitraire (par exemple, celle qui maximise la contribution au sein de la pioche). Notons en particulier que l'approche globale est plus rapide, mais est

moins précise, car elle pénalise toutes les arêtes simultanément. Dans les deux cas, le problème d'optimisation associé est le même. Le choix peut modifier en revanche la solution optimale.

2.6.b Normalisation de la matrice

La normalisation est très importante. En effet, en l'absence de contraintes de normalisation, la matrice identité est une structure d'appariement pour laquelle la variance est nulle (cf les remarques 35 et 35). La solution pour contraindre l'apprentissage et éviter de converger vers la structure "Identité" $\mathbb{W}[0]$ ou $\mathbb{B}[0]$, est d'imposer à chaque paire de séries de garder au minimum une connexion active. De plus, l'étape de normalisation de la matrice obtenue permet de concentrer sur les arêtes d'intérêt la pénalisation des arêtes.

La définition d'une structure de voisinage consiste à définir une distribution de probabilités sur les couples d'instants. En effet, l'objectif pour résoudre le problème d'optimisation booléen ou progressif est d'obtenir un système de pondération des liens entre tous les instants à travers une matrice de contiguïté. A cet effet, plusieurs choix de normalisation sont possibles. Les variantes de normalisation considérées dans le cadre de ce travail se divisent en deux familles. D'une part, le choix du niveau de la granularité (soit une normalisation par paires de séries, soit une normalisation toutes paires de séries confondues), et d'autre part, le choix de la temporalité (soit une normalisation par instants, soit une normalisation tous instants confondus).

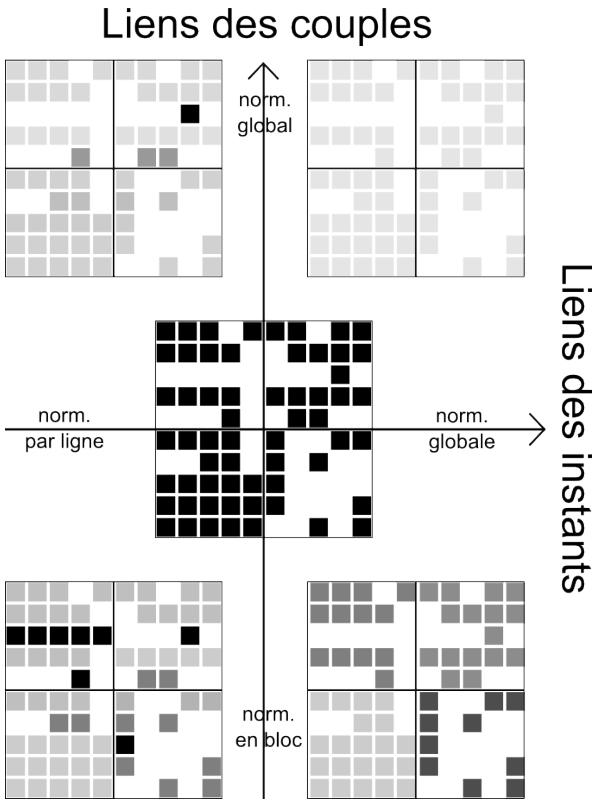


FIGURE 20 – Différents schémas de normalisation

La temporalité consiste à choisir de normaliser la matrice soit à l'échelle individuelle des instants, soit à l'échelle globale de l'ensemble des instants.

– La normalisation globale correspond à une distribution de la loi jointe des instants (en particulier, les deux séries comparées jouent un rôle symétrique)
– La normalisation par ligne correspond à une distribution de la loi conditionnelle des instants. Ainsi, à chaque instant est associée une distribution de probabilité (cette méthode est naturelle car la normalisation en ligne est une étape implicite lors du calcul de la variance, mais rompt la symétrie ligne/colonne du processus d'apprentissage.)

La granularité consiste à choisir de normaliser la matrice soit à l'échelle des paires de série, soit à l'échelle de l'ensemble des séries.

– La normalisation par paires assure un poids équitable à tous les couples de série dans la structure d'appariement.
– La normalisation globale, toutes paires de séries confondues permet d'extraire des couples de séries ayant un lien plus fort.

Ces deux types de variantes impactent la structure d'appariement finale et se traduisent donc par des contraintes supplémentaires dans le problème d'optimisation choisi.

2.7 Critère de fin

Le processus d'apprentissage est un processus itératif qui consiste à diminuer le poids des arêtes de l'ensemble pioche. Cependant, la contribution étant relative à la structure d'appariement, il y a toujours, à chaque itération, des arêtes dans l'ensemble pioche. La proposition suivante illustre cette affirmation.

Proposition 49 :

$$\begin{aligned} \forall\ (i,l) \in \{1,\dots,T\} \times \{1,\dots,n\}\ \ \{(i,l,i',l')/CM_{ii'}^{ll'} > 0\} = \varnothing \\ \iff\ \forall (i',l') \in \{1,\dots,T\} \times \{1,\dots,n\}\ \ CM_{ii'}^{ll'} = 0 \end{aligned}$$

Il faut donc définir un seuil d'arrêt pour ne pas poursuivre le processus à l'infini, et éviter que le processus ne modifie les poids d'arêtes intéressantes. Pour tenir compte de la spécificité des jeux d'observables, de la variabilité entre les différentes classes, et pour éviter d'aboutir à un problème de sur-apprentissage, il est préférable de définir certaines règles d'arrêt. Le fait de fixer arbitrairement le nombre d'itérations ne répond pas au problème.

2.7.a Définition de seuils d'arrêt fondés sur la variance

Critère de stabilité L'objectif de l'apprentissage est de trouver un appariement entre les séries, de sorte que la variance intra sous-jacente soit minimale. Nous pouvons définir un seuil s'arrêt, l'idée étant de continuer l'apprentissage jusqu'à ce que la chute de variance soit négligeable. On fixe alors un taux alpha fonction de la variance initiale, et dès que la chute globale de variance $V_{t+1} - V_t$ devient plus petite que ce taux, le processus d'apprentissage s'arrête.

Le critère d'arrêt consiste à sortir de la boucle si $\frac{V_{t+1}-V_t}{V_0} < \alpha$

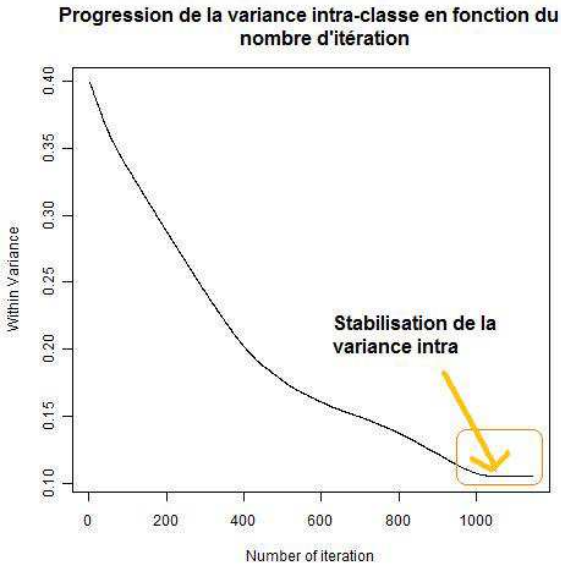
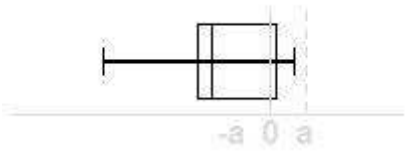


FIGURE 21 – Décroissance de la variance

2.7.b Définition de seuils d'arrêt fondés sur la contribution

Critère de tolérance En pratique, une arête peut appartenir à l'ensemble pioche malgré un faible impact sur la variance. Il est ainsi difficile de différencier une arête dont la contribution est faiblement négative d'une arête dont la contribution est faiblement positive. On fixe un seuil, en-dessous duquel la contribution d'une arête est considérée comme négligeable. Une arête, dont la contribution (en valeur absolue) ne dépasse pas le seuil, est exclue de l'ensemble pioche. On choisit de s'arrêter quand toutes les arêtes qui ne sont pas des arêtes d'intérêt ont une contribution négligeable. Le seuil peut être choisi en fonction de la variance ou de la contribution initiale.

Le critère d'arrêt consiste à sortir de la boucle si $\forall(i,j),\ C_{ij} < \alpha$, (cf figure 2.7.b)



Critère de neutralité Si au cours de l'apprentissage, toutes les arêtes appartiennent à la pioche, la pénalisation d'un ensemble d'arêtes revient à renforcer des arêtes qui doivent toutes être pénalisées. Plutôt que de choisir de privilégier une arête, le processus s'arrête.

Le critère d'arrêt consiste à sortir de la boucle si $\forall(i,j)C_{ij} > 0$. (cf. figure 22)

La figure 22 illustre les deux derniers critères. A la ligne 7, toutes les arêtes de poids non nul ont une contribution négative. Il n'y a pas d'arêtes dans la pioche. Dans les lignes 1,2 6 et 8, toutes les arêtes non nulles ont une contribution positive. Du fait du critère de neutralité,

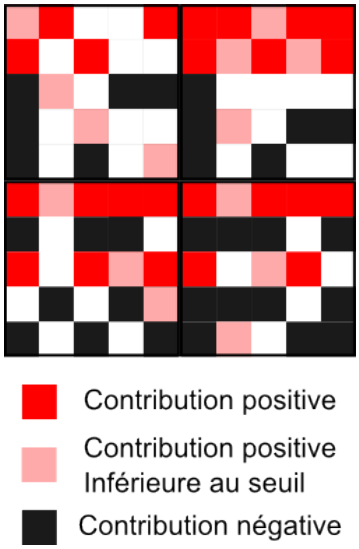


FIGURE 22 – Illustration des critères d'arrêt

l'apprentissage s'arrête. Finalement, les lignes 3, 4, 5, 9 et 10 ont des arêtes avec contribution positive (pioche non vide) mais dont la contribution est faible (critère de tolérance). Au regard des trois critères, le processus ayant donné cette matrice a abouti.

Cette section discute plusieurs variantes et aboutit à la définition de deux approches. En particulier, ces deux approches sont associées à des problèmes d'optimisation différents. Nous présentons donc dans une dernière section les deux approches retenues et nous discutons leur spécificité.

3 Discussion autour des différentes variantes

Nous avons, dans la section précédente, introduit plusieurs variantes. Les différents points introduits sont très dépendants de l'approche choisie et doivent répondre à la fois à des contraintes computationnelles et à des critères de qualité des appariements appris. Les deux problèmes d'optimisation évoqués, le problème booléen discret et le problème pondéré progressif, ont une finalité très différente. Les approches mises en œuvre varient. Nous ne considérons ici que l'approche intra-classe.

La première approche que nous considérons est l'approche booléenne.

3.1 Variante 1 : Cas d'une pénalisation booléenne ($\beta = 0$)

C'est le cas où la pénalisation consiste à annuler le poids des arêtes. L'idée sous-jacente est de rechercher parmi toutes les arêtes celles qui optimisent le critère, indépendamment de leur poids.

Convergence de cette approche Il est naturellement trivial, dans ce cas, que l'algorithme conduise à une diminution de la variance intra, la variation de la variance étant directement égale à la contribution. Annuler un poids est définitif et empêche tout retour ultérieur de l'arête dans le voisinage. Ainsi, le nombre d'itérations est borné.

Problème d'optimisation Dans le cadre d'un couplage initial équilibré, i.e., les arêtes initiales associées à un instant i sont toutes égales, l'approche booléenne permet de contraindre toutes les arêtes reliant un instant aux autres à être équi-pondérées. En effet, si un instant est pénalisé, son poids est redistribué équitablement entre toutes les autres arêtes de poids non nuls. Toutes les arêtes restantes conservent le même rapport de poids. Cette approche revient à chercher une solution au problème d'optimisation booléen.

$$\left\{\begin{array}{l} \text{Minimiser } V_W \text{ sous les conditions :} \\ \quad \forall k \in \{1,\ldots,K\},\ \forall (l,l') \in C_k,\ \forall (i,i') : \\ \qquad (i) w_{ii}^{ll} > 0 \text{ et } w_{ii'}^{ll} = 0 \text{ pour } i \neq i' \\ \qquad (ii) \exists (k_1,k_2)\ w_{k_1k_2}^{ll'} > 0 \\ \qquad (iii) \textbf{si } \mathbf{w_{ii'}^{ll'}} \neq \mathbf{0}, \mathbf{m_{ii'}^{ll'}} = \mathbf{m_{ii}^{ll}} \end{array}\right. \tag{42}$$

En particulier, cette formalisation rend le problème d'optimisation discret.

Nous présentons à présent l'approche progressive fondée sur des appariements non plus booléens, mais pondérés.

3.2 Variante 2 : Cas d'une pénalisation progressive

Une alternative à la pénalisation booléenne consiste à pénaliser fortement chaque arête, en lui laissant un poids faible, lui permettant d'être renforcé s'il y a lieu, en cours d'algorithme.

Convergence de cette approche Dans le cas d'une pénalisation forte ($\beta \approx 0$), la fonction Δ est de même signe qu'un polynôme en la variable β, elle est en particulier continue. La fonction ne change donc pas de signe dans un voisinage de 0. Pénaliser les arêtes dont la contribution est positive fait diminuer la variance ; à l'inverse, pénaliser les arêtes à contribution négative la fait augmenter.
Dans le cas d'une pénalisation faible ($\beta \approx 1$), nous étudions ici l'effet induit sur la variance par une pénalisation faible d'un poids choisi $m_{ii'}^{ll'}$ ((i,i',l,l') sont fixés). La variance dépendant de chacun des poids de l'ensemble $\left\{m_{ii'}^{ll'}\right\}$, nous pouvons en particulier la redéfinir comme une fonction V de $m_{ii'}^{ll'}$, vérifiant $V(m_{ii'}^{ll'}) = V_M$. Nous allons démontrer qu'une pénalisation faible d'une arête entraîne une variation de la variance de même signe que la contribution. Nous allons pour cela définir une fonction f de la manière suivante :

$$f : \beta \mapsto V_{m_{ii'}^{ll'}} - V(\beta\ m_{ii'}^{ll'}) \tag{43}$$

où $\beta \in [0,1]$ et $V(\beta\ m_{ii'}^{ll'})$ est la variance totale obtenue après pénalisation par un facteur β du lien (i,i',l,l') et renormalisation des liens $(i,i'',l,l')(i'' = 1,\cdots,T)$ pour satisfaire aux contraintes. f vérifie que $f(1) = 0$ et $f(0) = C_{ii'}^{ll'}$.

Nous introduisons deux autres fonctions δ_1 et δ_2 définies pour le triplet (i,j,l) par :

$$\begin{aligned} \delta_1(x_{i'j}^{l'}) &= x_{i'j}^{l'} - \sum_{r=1}^{M}\sum_{t=1}^{T} m_{it}^{lr} x_{tj}^{r} \\ \delta_2(i',l') &= \frac{m_{ii'}^{ll'}}{2(1-m_{ii'}^{ll'})}\left(1 + \frac{2m_{ii'}^{ll'}x_{i'j}^{l'}}{\delta_1(x_{ij}^{l}, x_{i'j}^{l'})}\right) \end{aligned}$$

La propriété suivante donne des conditions sous lesquelles le signe de la dérivée évaluée en 1 $f'(1)$ est différent du signe de la fonction en 0 $f(0)$.

Proposition 50 :

Soit Λ *le produit* $\delta_1(x_{ij}^l) \times delta_1(x_{i'j}^{l'})$. *Alors :*

1. $\text{signe}(-f'(1)) \neq \text{signe}(\Lambda) \Leftrightarrow 0 < \delta_1(x_{i'j}^{l'}) < m_{ii'}^{ll'} x_{i'j}^{l'}$
2. $\text{signe}(f(0)) \neq \text{signe}(\Lambda) \Leftrightarrow 0 < \delta_1(x_{ij}^l) < \delta_2(i', l')$ ou
   $0 < -\delta_1(x_{ij}^l) < -\delta_2(i', l')$

La preuve est mise en annexe (annexe B section 1). Ces cas sont proches du cas de convergence et arrivent rarement ; il suffit de faire un test au sein de l'algorithme pour les éviter.

Problème d'optimisation Dans le second cas, l'approche progressive permet de construire des voisinages pondérés, en modulant l'intensité des différentes arêtes. Un instant est pénalisé en fonction de sa contribution. Cette approche revient à chercher une solution au problème d'optimisation pondéré.

$$\begin{cases} \text{Minimiser } V_W \text{ sous les conditions :} \\ \quad \forall k \in \{1, \ldots, K\},\ \forall (l, l') \in C_k,\ \forall (i, i') \\ \quad\quad (i) w_{ii}^{ll} > 0 \text{ et } w_{ii'}^{ll} = 0 \text{ pour } i \neq i' \\ \quad\quad (ii) \exists (k_1, k_2) w_{k_1 k_2}^{ll'} > 0 \end{cases} \tag{44}$$

Cette approche vise à chercher une solution à un problème d'optimisation convexe. Elle admet donc un minimum global.

Nous allons maintenant comparer les deux approches.

3.3 Liens entre les différentes variantes

Les deux approches sont assez proches, mais certaines variantes de l'algorithme ne sont pas adaptées à l'une et à l'autre, et nous allons, dans cette section, présenter à partir des spécificités de chaque approche, les limites présentées par certaines variantes. Nous présenterons, pour terminer ce chapitre, une approche progressive et une approche booléenne permettant l'apprentissage des appariements discriminants.

3.3.a Différences conceptuelles entre les approches

Les diverses variantes évoquées consistent principalement à choisir une approche et un taux de pénalisation. Nous avons vu au paragraphe 3.2 que les différentes approches conduisent toutes en général à une variation de la variance allant dans le même sens que lors de la suppression totale de l'arête. L'approche qui semble la plus à même de conduire vers une structure optimale est l'approche locale et progressive avec une pénalisation faible. Cependant, la pénalisation faible d'une seule arête à chaque itération conduit à un processus très long, comme

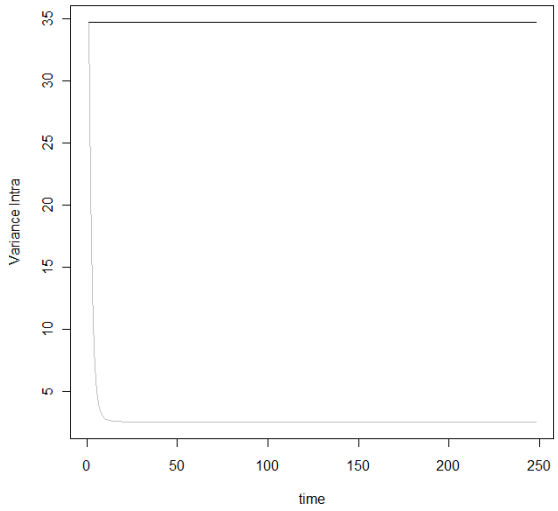


FIGURE 23 – Lenteur de l'algorithme pour des pénalisations faibles

nous pouvons l'observer sur la figure 23. Tandis que l'approche correspondant à une pénalisation forte des arêtes a convergé (courbe grise), la variance pour l'approche fondée sur des pénalisations faibles n'a presque pas été diminuée (courbe noire).

Pour accélérer le processus, se présentent donc deux options, le choix d'une approche globale, consistant à augmenter le nombre de modifications simultanées ou le choix d'une pénalisation forte des arêtes. Nous avons vu précédemment que ces deux approches conduisent à des problèmes d'optimisation complètement différents.

- **Problèmes liés à l'approche globale** L'approche globale ne tient pas compte des liens existant entre instants. Le calcul de la contribution est fait sur la base d'une unique arête pénalisée, tandis que l'approche globale considère plusieurs pénalisations. De plus, la pénalisation d'une arête de la pioche a une incidence sur les contributions des autres, et donc, pénaliser une arête contribuant fortement à la variance peut modifier l'ampleur et le signe des contributions d'autres arêtes à l'itération suivante.
- **Problèmes liés aux pénalisations fortes** Dans le cas où une pénalisation forte des arêtes de la pioche est effectuée, le fait de ne pas mettre à 0 le poids conduit à un ensemble de poids faibles, qui peuvent induire du bruit au sein de l'analyse. Ceci se traduit par un ralentissement de l'algorithme, et un reliquat d'arêtes dont le poids est faible, mais pas nul. De plus, lorsque la pénalisation initiale est trop forte, les arêtes ne peuvent pas "revenir" facilement. Le cas extrême est l'approche définitive qui consiste à mettre à 0 le poids des arêtes à pénaliser. Cette approche vide la structure d'arêtes très faibles mais empêche une arête pénalisée à un instant donné, en raison d'un voisinage initial chaotique, de voir son poids ré-augmenter, dans le cas d'un voisinage resserré sur les arêtes d'intérêt.

Bilan Pour résumer, les deux choix qui se présentent pour l'intensité de la pénalisation découlent de paradigmes fondamentalement différents. Une pénalisation booléenne vise à avoir une structure plus vidée, avec une distinction entre les arêtes pénalisées et les arêtes

conservées, mais pas de distinction au sein des poids des arêtes d'intérêt. Une pénalisation progressive, forte ou faible, conduit au contraire à une structure dérivée du couplage complet, où existe une variabilité des poids, au sein même de l'ensemble d'intérêt. Les arêtes gardent toutes un poids latent, parfois très faible. L'intensité des pénalisations aura une incidence sur la qualité et la rapidité de convergence. Les deux approches induisent des minimums locaux. Les deux méthodes que nous proposons induisent une diminution de la variance, mais n'assurent pas d'obtenir une variance minimale. Cependant, une pénalisation forte a plus de risque de conduire à une mauvaise configuration, car l'arête n'a plus l'opportunité de voir son poids augmenter. Quant à la différence entre l'approche locale et l'approche globale, le choix de l'une ou de l'autre conduit à des résultats très différents. Par l'approche globale, les poids sont modifiés simultanément, en dépit de liens sous-jacents. En effet, la modification d'une arête à un instant t a une incidence sur les contributions d'autres arêtes à l'instant t+1. Cependant, le fait de ne pénaliser qu'une seule arête fortement peut induire un déséquilibre entre des arêtes qui se ressemblent fortement. Notons que l'approche locale peut s'effectuer selon une règle séquentielle, ou parallèle par blocs, avec une incidence sur le résultat final

3.3.b Stratégie adoptée

Dans la suite, nous considérerons les deux approches : l'approche booléenne locale et l'approche progressive globale avec faible pénalisation. Ces deux approches donnent des voisinages minimisant tous les deux la variance intra (respectivement maximisant la variance inter) en répondant à deux problèmes d'optimisation différents.

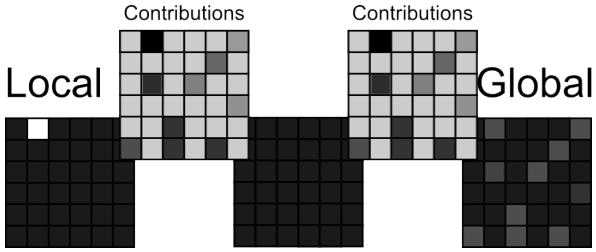


FIGURE 24 – Différences entre les approches locales et globales

3.3.c Problèmes de normalisation

Résumons en termes d'optimisation, les différences apportées par les variantes définies à la section 2.6.b : En ce qui concerne la temporalité, la normalisation en ligne assure le fait qu'un lien soit préservé pour chaque instant et que chaque instant conserve un poids équivalent. Au contraire, une normalisation globale peut privilégier certains instants.
En ce qui concerne la granularité, la normalisation par blocs donne des poids équivalents à toutes les paires de séries, tandis que la normalisation globale permet de moduler l'importance des séries. Cependant, la normalisation globale ne contraint plus l'optimisation à éviter les cas limites précédents.

Normalisation en ligne des blocs La normalisation en ligne impose pour chaque instant i de la série S^l que les poids des arêtes croisant l'instant i avec tous les instants de $S^{l'}$

définissent une distribution de probabilité. Le bloc décrit la distribution des poids des arêtes du bloc conditionnellement à chaque arête.

Cela se traduit par la condition (ii) dans le problème d'optimisation ci-dessous, qui assure de plus la non-nullité des blocs.

$$\begin{cases} \text{Minimiser } V_W \text{ sous les conditions :} \\ \quad \forall k \in \{1,\dots,K\},\ \forall (l) \in C_k,\ \forall (i) \\ \quad\quad (i) \forall (i') w_{ii}^{ll} > 0 \text{ et } w_{ii'}^{ll} = 0 \text{ pour } i \neq i' \\ \quad\quad (ii) \forall (\mathbf{l'}) \sum_{\mathbf{i'=1}}^{\mathbf{T}} \mathbf{w_{ii'}^{ll'}} = \mathbf{1} \end{cases} \tag{45}$$

Comparaison des deux approches de normalisation La normalisation globale des blocs donne plus de force à des instants pour lesquels le voisinage est dense. Plus on conserve d'arêtes reliant deux instants, plus le poids de l'instant est important. Au contraire, la normalisation en ligne donne un poids équivalent à tous les instants, en dépit de leur caractère discriminant.

Le choix de la normalisation dépend de l'approche choisie.

1. Dans le cadre d'une approche progressive globale, chaque bloc étant pénalisé simultanément, la normalisation est importante. Pour respecter une équi-pondération entre les instants des séries, la normalisation en ligne des appariements entre paires de séries sera privilégiée dans le cadre de cette approche.
2. Dans le cas de la pénalisation booléenne locale, les deux types de normalisation ne sont pas adaptées. En effet, à chaque itération, une seule arête $m_{ii'}^{ll'}$ est pénalisée. Ainsi, les autres arêtes liant l'instant i voient leur poids augmenter, tendant donc à majorer leur contribution. Ainsi, le processus tend à pénaliser davantage les arêtes liant des instants ayant déjà été pénalisés. Pour éviter ces inconvénients, nous allons choisir une normalisation en ligne, toutes séries confondues, et nous allons, dans le cadre de cette approche, définir des critères d'arrêt pour éviter la convergence de l'algorithme vers la structure "Identité".

La solution adoptée consiste à **imposer certaines contraintes et condition d'arrêt** dans le processus d'apprentissage, évitant la convergence vers la structure d'appariement triviale Identité. Nous détaillons ce second point dans la section 3.3.d. La normalisation globale impose, au sein de chaque ensemble de séries (à l'échelle d'un couple), d'avoir une distribution de probabilité associée à l'ensemble des arêtes. Le bloc décrit la distribution jointe des poids des arêtes. Cela se traduit par la condition (iii) dans le problème d'optimisation ci-dessous.

$$\begin{cases} \text{Minimiser } V_W \text{ sous les conditions :} \\ \quad \forall k \in \{1,\dots,K\},\ \forall (l,l') \in C_k,\ \forall (i,i') \\ \quad\quad (i) w_{ii}^{ll} > 0 \text{ et } w_{ii'}^{ll} = 0 \text{ pour } i \neq i' \\ \quad\quad (ii) \exists (k_1,k_2) w_{k_1k_2}^{ll'} > 0 \\ \quad\quad (iii) \sum_{\mathbf{k_1=1}}^{\mathbf{T}} \sum_{\mathbf{k_2=1}}^{\mathbf{T}} \mathbf{w_{k_1k_2}^{ll'}} = \mathbf{1} \end{cases} \tag{46}$$

3.3.d Sélection d'arêtes et conditions d'arrêt, critère de définition de la pioche

Les deux approches que nous adoptons consistent à définir les arêtes de la pioche comme les arêtes qui ne sont pas d'intérêt, celles qui ont tendance à faire augmenter la variance lorsqu'on cherche à la minimiser, et celles qui ont tendance à faire diminuer la variance lorsqu'on cherche à la maximiser. Dans le processus de sélection des arêtes, à chaque itération, nous trouvons des arêtes dont la contribution est positive.
En particulier, dans le cadre d'une pénalisation progressive, si on choisit les arêtes en fonction du signe de leur contribution : l'algorithme consistant à pénaliser certaines arêtes de la pioche est ainsi un processus infini. Dans le cadre de l'approche booléenne, nous avons vu que l'apprentissage conduit à un appariement Identité. De plus, dans les deux cas, nous distinguons dans l'algorithme deux phases, la première où sont pénalisées les arêtes apportant beaucoup de variabilité, puis la seconde où sont pénalisées les arêtes dont la contribution a changé de signe en cours d'algorithme. Ainsi, nous voyons apparaître un phénomène de sur-apprentissage. Il faut donc définir des conditions d'arrêt.
Dans le cadre de l'approche progressive, aucune arête n'est jamais tuée. Au fil des itérations, les poids peuvent diminuer jusqu'à devenir négligeables. Les seuils d'arrêt, fondés sur la variance, que nous évoquons dans le paragraphe 2.7.a permettent de stopper le processus d'apprentissage lorsque l'algorithme converge vers la structure recherchée : ce problème s'apparente à un problème de seuillage. Sans ces critères, l'algorithme ne s'arrête jamais, la modification de la variance devient négligeable à chaque itération et des arêtes de poids faible continuent à être pénalisées. Dans la version progressive de l'algorithme, nous optons donc pour ce type de contraintes qui nous assurent la stabilité de l'algorithme.
Dans la forme booléenne de l'algorithme, le nombre d'arêtes diminue nécessairement à chaque itération, et la méthode se termine nécessairement. L'objectif n'est donc plus d'assurer la convergence et l'arrêt de l'algorithme, mais d'imposer des contraintes sur les poids pour éviter le sur-apprentissage (structure "Identité" ou autre structure creuse). Il faut éviter de pénaliser les arêtes ayant des effets négligeables sur la variance et il faut choisir au sein d'arêtes équivalentes de la pioche certaines arêtes plutôt que d'autres. Nous introduisons à ce propos des critères de neutralité et de tolérance.
Nous limitons d'une part la définition des arêtes appartenant à la pioche à celles dont la contribution dépasse un certain seuil (critère de tolérance). Nous avons choisi, dans le cadre de notre apprentissage, de prendre pour seuil un certain pourcentage de la contribution maximale associée à une distribution uniforme. D'autre part, pour éviter de converger vers une structure vide, et imposer de garder pour chaque couple $(S^l, S^{l'})$ et chaque instant de S^l au moins un lien vers les instants de $S_{l'}$, nous ne pénalisons pas les instants d'une série au sein d'un couple où aucune arête n'est arête d'intérêt (critère de neutralité). Nous redéfinissons donc la pioche pour introduire ces deux concepts.

Définition 51 : (Pioche Π dans le cadre intra-classe)

$$\Pi = \left\{ (i,i',l,l') \backslash \begin{array}{r} C[W]^{ll'}_{ii'} > \alpha max(|C[J]^{l_1 l_2}_{i_1 i_2}|) \\ \exists (i'',l'') C[W]^{ll'}_{ii'} < 0 \end{array} \right\} \quad (47)$$

Définition 52 : (Pioche Π dans le cadre inter-classes)

$$\Pi = \left\{ (i,i',l,l') \backslash \begin{array}{r} C[W]_{ii'}^{ll'} < \alpha max(|C[J]_{i_1 i_2}^{l_1 l_2}|) \\ \exists (i'',l'') C[W]_{ii'}^{ll'} > 0 \end{array} \right\} \quad (48)$$

Ces contraintes se traduisent dans le problème d'optimisation de la manière suivante :

$$\left\{ \begin{array}{l} \text{Minimiser } V_W \text{ sous les conditions :} \\ \quad \forall k \in \{1,\ldots,K\},\ \forall (l) \in C_k,\ \forall (i) \\ \qquad (i) \forall (i') w_{ii}^{ll} > 0 \text{ et } w_{ii'}^{ll} = 0 \text{ pour } i \neq i' \\ \qquad (ii)\ \sum_{l' \in C_k} \sum_{i'=1}^{T} w_{ii'}^{ll'} = 1 \\ \qquad (iii) \forall (\mathbf{i'},\mathbf{l'}) \mathbf{w_{ii'}^{ll'}} \notin \mathbf{\Pi} \end{array} \right. \quad (49)$$

3.4 Initialisation de la matrice de poids

Le choix d'une matrice pour l'apprentissage de l'algorithme est un élément fondamental de l'approche afin de combiner les apprentissages intra et inter en vue de sortir des appariements discriminants entre paires de série. Ce point sera donc traité plus en détail dans le chapitre suivant.

Conclusion

Structure apprise Nous avons proposé dans ce chapitre deux manières d'apprendre des appariements. Les deux approches sont communes dans la forme, elles consistent toutes deux, par un processus itératif, à rechercher au sein de l'ensemble des arêtes, celles qui entraînent une augmentation de la variance intra ou une diminution de la variance inter. Ces arêtes sont alors pénalisées jusqu'à la convergence de l'algorithme. En revanche, les deux variantes que nous avons proposées ont un sens très différent. L'approche progressive apprend des appariements proches d'un couplage complet. Les arêtes obtenues ont un poids dont l'intensité dépend du caractère discriminant de l'arête. L'approche booléenne consiste à apprendre des appariements indépendamment d'un système de pondération. L'objectif consiste à trouver au sein de l'ensemble des arêtes initiales les arêtes d'intérêt, sans considération de leur poids. La première approche apprend un système de poids, tandis que la seconde apprend un ensemble d'arêtes.

Etude des problèmes d'optimisation sous-jacents Nous cherchons une structure d'appariement M* vérifiant certaines contraintes. Nous avons donc, dans ce qui précède, proposé deux problèmes d'optimisation et leur réalisation algorithmique pour apprendre de tels appariements, en vue de minimiser la variance intra et de maximiser la variance inter.

3.5 Approche progressive

$$\begin{cases} \text{Minimiser } V_W \text{ sous les conditions :} \\ \quad \forall k \in \{1,\ldots,K\},\ \forall (l,l') \in C_k,\ \forall (i,i') \\ \qquad (i) w_{ii}^{ll} > 0 \text{ et } w_{ii'}^{ll} = 0 \text{ pour } i \neq i' \\ \qquad (ii) \sum_{k=1}^{T} w_{ik}^{ll'} = 1 \end{cases} \tag{50}$$

Proposition 53 :

Le problème de minimisation est un problème convexe.

Le problème étant un problème convexe, il admet donc un minimum global sur l'espace défini par les contraintes. Pour trouver une solution convergeant vers le minimum global, nous pourrions employer d'autres méthodes, à l'instar de la méthode du gradient projeté. Nous comparerons notre approche à cette dernière dans le chapitre suivant.

3.6 Approche booléenne

$$\begin{cases} \text{Minimiser } V_W \text{ sous les conditions :} \\ \quad \forall k \in \{1,\ldots,K\},\ \forall (l,l') \in C_k,\ \forall (i,i') \\ \qquad (i) w_{ii}^{ll} > 0 \text{ et } w_{ii'}^{ll} = 0 \text{ pour } i \neq i' \\ \qquad (ii) \exists (k_1,k_2) w_{k_1k_2}^{ll'} > 0, \\ \qquad (iii) \mathbf{Siw_{ii'}^{ll'} \neq 0}, \mathbf{m_{ii'}^{ll'} = m_{ii}^{ll}} \end{cases} \tag{51}$$

Proposition 54 :

Le problème de minimisation est un problème discret.

La recherche exhaustive est la seule méthode d'optimisation donnant un optimum global.

Chapitre 4

Apprentissage d'appariements discriminants : mise en œuvre

Il est fondamental pour l'exploration et la discrimination de séries temporelles d'apprendre des appariements qui soient discriminants. Nous avons présenté un problème d'optimisation visant à cet apprentissage. Nous présentons ici les détails algorithmiques du processus d'apprentissage. Nous combinons dans un premier temps l'algorithme apprenant des appariements caractérisant les liens entre séries au sein d'une classe et l'algorithme apprenant des appariements différentiant les séries de classes différentes. Nous comparons la complexité calculatoire de notre approche aux méthodes d'optimisation fondés sur les méthodes de gradient et étudions les appariements appris

Le chapitre précédent a introduit les concepts théoriques liés au problème d'optimisation et les contraintes algorithmiques associées. Nous présentons dans cette partie la mise en œuvre pratique de l'apprentissage d'appariements en vue de la discrimination d'un ensemble de séries temporelles. Dans un premier temps, nous présentons l'aspect algorithmique, introduisant les conditions initiales et la façon de coupler les approches caractéristiques et différentielles en vue de l'apprentissage d'un couplage discriminant. Nous présentons ensuite le détail de l'algorithme, ainsi que sa complexité. Finalement, nous observons certains résultats sur des jeux de données simulés.

1 Spécificité et combinaison de l'apprentissage de structures caractéristiques et différentielles

Nous avons montré dans les chapitres précédents l'importance de la réunion de l'apprentissage intra et inter-classes en vue de la discrimination d'une partition de séries temporelles. Nous avons également discuté de la difficulté de cette réunion. Nous présentons ici quelques pistes pour accomplir cette tâche à travers l'initialisation des matrices d'appariement.

1.1 Définition d'un critère de sélection spécifique à chaque approche

Nous avons opté pour une approche d'apprentissage qui repose sur la définition de trois objets :

1. un critère de discrimination permettant de définir la "pioche" (i.e., l'ensemble des arêtes à pénaliser)
2. un procédé de sélection au sein de la pioche et des critères d'arrêt
3. la donnée d'un ensemble d'arêtes initiales,

Les arêtes sélectionnées dans la pioche sont pénalisées de manière itérative, et leur poids est redistribué à d'autres arêtes. En ce qui concerne le critère de discrimination (point 1), les deux approches que nous proposons reposent sur le même critère. Une arête est discriminante si elle induit une diminution de la variance intra (moindre variabilité au sein des classes) et une augmentation de la variance inter (variabilité maximale entre les classes). Pour la sélection des arêtes (point 2), les définitions de la pioche et des critères d'arrêt diffèrent selon la méthode choisie. La première approche consiste à supprimer à chaque itération une arête dominante au sein de la pioche (approche booléenne locale) selon des critères de tolérance et de neutralité, la seconde consiste à pénaliser faiblement toutes les arêtes de la pioche (approche progressive globale) jusqu'à un critère de stabilité. En ce qui concerne les arêtes initiales (point 3), le processus d'apprentissage des appariements au sein des classes et entre les classes vise à apprendre séparément des appariements caractéristiques et des appariements différentiels. Cette différenciation entre le caractéristique et le différentiel conduit à séparer les couplages en fonction de la nature du lien. Le fait de travailler avec des arêtes de couples et non pas des arêtes sémantiques, implique la différentiation de deux types d'arêtes :

- les arêtes liant deux séries d'une même classe
- les arêtes liant deux séries de deux classes différentes

Pour coupler les deux approches, nous proposons d'initialiser l'apprentissage des arêtes caractéristiques (respectivement différentielles) à l'aide d'une structure de voisinage fondée sur les matrices apprises en sortie de l'apprentissage différentiel (respectivement caractéristique).

1.2 Différenciation des algorithmes intra et inter

Les algorithmes d'apprentissage intra-classe et inter-classes sont séparés de manière naturelle, par la nature distincte des deux types de matrice d'appariement. Dans les deux cas, le processus d'apprentissage des appariements au sein des classes et entre les classes, vise à apprendre séparément des appariements caractéristiques et des appariements différentiels. Cette distinction entre le caractéristique et le différentiel oblige à séparer les couplages en fonction de la nature du lien. Il semble cependant plus intéressant de trouver les arêtes optimisant un critère de discrimination. La solution choisie pour coupler les deux types d'apprentissage, consiste à initialiser le processus par une structure fondée sur les sorties précédentes :

- les arêtes faisant partie d'un bloc intra, choisies pour leur potentiel à homogénéiser la structure de voisinage, sont sélectionnées parmi des couples d'instants différenciant les classes.
- les arêtes faisant partie d'un bloc inter, choisies pour augmenter l'hétérogénéité des classes, sont sélectionnées parmi les couples d'instants caractérisant leur classe.

En particulier, les arêtes discriminantes sont les arêtes intra caractéristiques parmi celles qui différencient les classes, et les arêtes inter différentielles parmi celles qui caractérisent les classes.
Nous souhaitons à présent lier les approches intra et inter par l'initialisation des matrices d'appariement. Nous proposons, dans la prochaine sous-section, d'étudier deux variantes pour cette étape.

1.3 Séparation des classes

Une différence fondamentale entre les deux algorithmes intra et inter réside dans le fait que l'apprentissage intra vise à étudier les appariements entre toutes les paires de séries au sein d'une classe, tandis que l'apprentissage inter peut concerner plusieurs classes. L'apprentissage inter peut ainsi se faire par paires de classes, ou bien toutes classes confondues.
La première approche cherche à différentier une série avec chaque classe prise individuellement : cette approche permet de trouver les instants qui différencient deux séries appartenant à des classes très proches.
La seconde approche cherche à différencier une série avec toutes les autres classes prises simultanément : cette approche permet de souligner les arêtes globalement différentielles, mais ne fait pas la distinction entre des classes de séries proches et des classes de séries éloignées.

Apprentissage inter croisant une classe avec toutes les autres Cette approche consiste à rechercher les événements globalement différentiels vis-à-vis des autres classes. L'apprentissage pour la classe *cl* conduit alors à initialiser B par des blocs propres à chaque série et constants au sein de toutes les classes :

$$B_{ll'} = \begin{cases} I_T & \text{si } y_l = y_{l'} = cl \\ \text{seuil}(\bar{W}^{l.}) & \text{si } y_{l'} \neq y_l \end{cases} \tag{52}$$

Apprentissage par paires de classe Cette approche consiste à sélectionner des événements différentiels pour chaque couple de classes. Ceci nécessite un apprentissage séparé pour toutes les classes. L'apprentissage pour la classe *cl* conduit alors à initialiser B par des blocs : nuls partout sauf pour les blocs croisant la classe cl à la classe cl'.

$$B_{ll'} = \begin{cases} I_T & \text{si } y_l = cl \text{ et } l = l' \\ \text{seuil}(\bar{W}^{l.}) & \text{si } \mathbf{cl'} = \mathbf{y_{l'}} \neq y_l \end{cases} \tag{53}$$

Choix de l'approche par paires de classe L'apprentissage par paires de classe est une solution plus spécifique. Elle permet de limiter l'algorithme à la distinction de classes proches, ce qui constitue le cœur de notre approche ; dans la suite, l'apprentissage de la structure de voisinage inter-classes se fera donc entre paires de classes. A partir de cette dernière étape de fusion des deux algorithmes intra et inter, nous pouvons à présent apprendre des appariements qui soient discriminants. Nous présentons dans la suite un détail des procédés algorithmiques mis en œuvre pour l'apprentissage des appariements.

1.4 Mise en œuvre de l'apprentissage discriminant

Nous proposons ici le détail des algorithmes LearnW et LearnB qui gèrent l'apprentissage des appariements en fonction de la structure initiale. Nous avons vu que l'approche d'ap-

prentissage que nous définissons repose sur trois points, dont l'introduction d'un critère de discrimination permettant de définir la "pioche" (i.e., l'ensemble des arêtes à pénaliser), un procédé de sélection au sein de la pioche et des critères d'arrêt, et encore la donnée d'un ensemble d'arêtes initiales. Nous présentons tout d'abord le détail algorithmique de la définition de la pioche.

1.4.a Définition de la pioche et de l'ensemble des arêtes d'intérêt

Algorithme intra-classe LearnW L'algorithme LearnW a pour objectif de trouver les arêtes minimisant la variance intra. Les arêtes d'intérêt sont constituées des arêtes dont la contribution à la variance CW_{ij} est négative, c'est-à-dire dont la suppression et la redistribution des poids entraîne une hausse de la variance. La pioche est construite à partir des arêtes dont la contribution est positive, et au-dessus du seuil de tolérance α_W Notons que ce pourcentage α_W est fonction de la variabilité des jeux et est à déterminer pour chaque jeu de données.

Définition 55 : (Ensemble des arêtes d'intérêt E_W)

$$E_W = \left\{(i,j,l,l')\backslash CW_{ij}^{ll'} < 0\right\} \tag{54}$$

Définition 56 : (Pioche Π)

$$\Pi = \left\{ (i,j,l,l')\backslash \left\{ \begin{array}{c} CW_{ij}^{ll'} > \alpha \ \max(|CW_{ij}^{ll'}|) \\ \exists j \backslash (i,j,l,l') \in E_W \end{array} \right\} \right. \tag{55}$$

Algorithme inter-classes LearnB Par symétrie, l'algorithme LearnB a pour objectif de trouver les arêtes maximisant la variance inter. Les arêtes d'intérêt sont constituées d'arêtes dont la contribution à la variance CB_{ij} est positive, c'est-à-dire dont la suppression et la redistribution des poids entraîne une diminution de la variance. La pioche est construite à partir des arêtes dont la contribution est négative, et inférieure au seuil de tolérance α_B (à nouveau propre à chaque jeu d'observables).

Définition 57 : (Ensemble des arêtes d'intérêt E_B)

$$E_B = \left\{(i,j,l,l')\backslash CB_{ij}^{ll'} < 0\right\} \tag{56}$$

Définition 58 : (Pioche Π)

$$\Pi = \left\{ (i,j,l,l') \backslash \left\{ \begin{array}{r} CB_{ij}^{ll'} < -\alpha \ \max(|CB_{ij}^{ll'}|) \\ \exists j \backslash (i,j,l,l') \in E_B \end{array} \right. \right\} \tag{57}$$

Nous avons donc présenté la définition de la pioche. Nous avons dans ce qui précède montré comment l'initialisation de la matrice jouait un rôle fondamental pour apprendre des appariements qui soient discriminants. Nous présentons ici la mise en œuvre de cette initialisation.

1.4.b Initialisation de la matrice de poids

L'approche d'apprentissage repose également sur la donnée d'un ensemble d'arêtes initial. Pour notre objectif, consistant à apprendre des appariements discriminants, nous voudrions combiner les deux approches. Or un appariement entre deux séries est un couplage qui sera d'une part caractéristique vis-à-vis des séries de sa classe, et différentiel vis-à-vis des séries des autres classes.
Pour coupler les deux approches, nous avons choisi de chercher les arêtes caractéristiques (respectivement différentielles) uniquement au sein des arêtes qui sont potentiellement différentielles (respectivement caractéristiques). Cela revient dans le processus intra (respectivement inter) à initialiser la structure de voisinage à partir des arêtes différentielles (respectivement caractéristiques), que nous pouvons définir à partir de la sortie du processus inter (respectivement intra). Les deux algorithmes d'apprentissage restent séparés, mais nous proposons une approche séquentielle permettant de les coupler.

Initialisation de la matrice Intra Notons B une structure de voisinage. Chaque série S_l de la classe y_l est associée avec chaque série $S_{l'}$ dans une classe $y_{l'} \neq y_l$ selon une structure de voisinage $B^{ll'}$. Définissons $\bar{B}^l$, le bloc moyen associé aux blocs inter associés à la série S^l.

$$\bar{B}^l = \frac{1}{n - n_l} \sum_{y_{l'} \neq y_l} B^{ll'} \tag{58}$$

Utilisons les $\bar{B}$ pour définir le bloc $W_{ll'}$ initialisant le processus d'apprentissage intra.

Cas de l'approche booléenne Pour les problèmes de proportionnalité évoqués au paragraphe 2.4.b du chapitre 3 de cette partie, il est préférable d'éviter l'initialisation du processus d'apprentissage par une matrice pondérée. En effet, cela a pour effet de diminuer l'impact des arêtes dont le poids est faible, et de pénaliser en priorité les arêtes dont le poids est fort. Cela revient à se placer dans l'espace défini par les contraintes du problème d'optimisation booléen associé. Nous adoptons donc un seuillage de la matrice $\bar{B}^l$. Une arête se dégage du processus d'apprentissage si l'arête est majoritairement sélectionnée dans B. Ainsi, si le poids d'une arête sémantique $\bar{B}^l_{ii'}$ est plus faible que le poids d'un couplage complet, nous initialisons cette arête dans chaque bloc de la structure intra à apprendre, avec un poids nul.

$$W_{ll'} = \begin{cases} \frac{1}{n_{cl_1}} I_T & \text{si } l = l' \\ \frac{1}{n_{cl_1} \text{seuil}(\bar{B}^l)} & \text{si } y_l = y_{l'} \end{cases} \tag{59}$$

avec

$$\textit{seuil} : \mathbb{M}_{\mathbb{T}}(\mathbb{R}) \to \mathbb{M}_{\mathbb{T}}(\mathbb{R}) \text{ tel que } \begin{cases} \text{seuil}(B)_{ij} = \frac{1}{\#\{B_{ij} > \frac{1}{T}\}} & \text{si } B_{ij} > \frac{1}{T} \\ \text{seuil}(B)_{ij} = 0 & \text{sinon} \end{cases} \tag{60}$$

En particulier, quand le voisinage inter est $\mathbb{B}[J]$ (i.e., aucune arête n'a été pénalisée), le voisinage intra correspondant est $\mathbb{W}[J]$.

Cas de l'approche progressive Dans ce cas, il n'y a pas de problèmes de proportionnalité. En particulier, la matrice $\bar{B}^l$ vérifie les contraintes du problème d'optimisation associé. La normalisation en ligne des blocs initiaux est respectée.

$$W_{ll'} = \begin{cases} I_T & \text{si } l = l' \\ \bar{B}^l & \text{si } y_l = y_{l'} \end{cases} \tag{61}$$

En particulier, quand le voisinage inter est $\mathbb{B}[J]$ (i.e., aucune arête n'a été pénalisée), le voisinage intra correspondant est à nouveau $\mathbb{W}[J]$.

Initialisation de la matrice Inter Notons W une structure de voisinage. Chaque série S_l de la classe y_l est associée aux séries S de la classe y_l selon une structure de voisinage $W^{ll'}$. Par symétrie, définissons $\bar{W}^l = \frac{1}{n_l - 1} \sum_{\substack{y_{l'} = y_l \\ l \neq l'}} B^{ll'}$. Utilisons les blocs sémantiques $\bar{W}^l$ pour définir le bloc $B_{ll'}$ initialisant le processus d'apprentissage inter.

Cas de l'approche booléenne En adoptant le seuillage de la matrice $\bar{W}^l$, nous obtenons la matrice d'initialisation inter

$$B_{ll'} = \begin{cases} I_T & \text{si } l = l' \\ \text{seuil}(\bar{W}^l) & \text{si } y_l \neq y_{l'} \end{cases} \tag{62}$$

En particulier, quand le voisinage intra est $\mathbb{W}[J]$ (i.e., aucune arête n'a été pénalisée), le voisinage inter correspondant est $\mathbb{B}[J]$.

Cas de l'approche progressive A nouveau, pas de seuillage dans ce cas,

$$B_{ll'} = \begin{cases} I_T & \text{if } l = l' \\ \bar{W}^l & \text{si } y_l \neq y_{l'} \end{cases} \tag{63}$$

Cette section a permis de présenter la manière dont l'algorithme discriminant était mis en œuvre, à partir des deux algorithmes LearnW et LearnB. Nous allons à présent dans la section qui suit présenter sommairement les grandes étapes de l'algorithme

2 Raffinage de l'algorithme

Nous présentons dans cette section le détail de l'algorithme proposé pour l'apprentissage des poids discriminants. Nous détaillons dans un premier temps les deux approches intra et inter-classes, puis nous introduisons le processus de combinaison des deux algorithmes.

2.1 Structure globale de l'algorithme intra-classe

Rappelons la structure générale des deux algorithmes.

Algorithm 2 Structure générale de l'algorithme

1: Initialisation : définition de la matrice initiale
2: **repeat**
3: Définition du critère de sélection des arêtes
4: Mise à jour de la matrice de voisinage
5: **until** critère de fin
6: **return** La matrice de voisinage obtenue

Nous allons, dans cette section, rentrer de façon sommaire dans le détail de chaque étape au sein des deux algorithmes intra et inter. Nous présentons tout d'abord l'algorithme d'apprentissage de la matrice d'appariement intra.

2.2 Apprentissage des appariements intra

Nous présentons ici toutes les étapes de l'algorithme intra-classe.

Initialisation des matrices ˜
Cet algorithme d'initialisation prend comme paramètre d'entrée une structure de voisinage inter. Il renvoie en sortie la matrice W initiale. W est une structure de voisinage intra.

Algorithm 3 initialisation

1: Entrée : $Binit_{(nT\times nT)}$
2: Sortie : $W_{(nT\times nT)}$
3: **for** classe cl, séries $l, l' \in [\![1,n]\!]\times[\![1,n_{cl}]\!]^2$ **do**
4:

$$W^{ll'} = \frac{1}{n-n_{cl}} \sum_{\substack{k\in 1..n \\ y_k \neq y_l}} Binit^{lk}$$

5: **end for**

L'initialisation se fait à partir de la structure de voisinage inter en cours d'apprentissage. Une moyenne des blocs inter correspondant à cette série S^l est effectuée, et définit le bloc intra initial choisi pour cette série pour initialiser tous les blocs $W^{ll'}$, de sorte que les contraintes de normalisation soient respectées. A partir de cette matrice initiale, nous initions le processus de sélection des arêtes.

Définition du critère de sélection des arêtes L'algorithme de sélection des arêtes prend comme paramètre d'entrée une structure de voisinage intra, ainsi que la matrice des valeurs $X = (X_1, .., X_p)$, où $X_k = (x^1_{1k}..x^1_{Tk}, x^2_{1k}.., x^n_{Tk})$. Il renvoie la pioche et la matrice des

contributions C. Les contributions sont calculées selon la formule des centres mobiles. Cette procédure parcourt l'ensemble des couples d'instants (i, i') de chaque paire de séries (l, l').

Algorithm 4 select-aretes

1: Entrée : structure de voisinage intra $W_{(nT\times nT)}$, données $X_{(nT\times p)} = (X_1, .., X_p)$, classe cl, seuil θ
2: Sortie : Pioche, matrice des contributions $CW_{(n_{cl}T\times n_{cl}T)}$
3: Pioche $= \varnothing$
4: **for** (serie l,l', instants i,i') $\in [\![1, n_{cl}]\!]^2\times[\![1, T]\!]^2$ **do**
5: Calculer $CW_{i,i',l,l'}$ (formule des centres mobiles)
6: Ajouter à l'ensemble Pioche : $\{(i, i', serie^l, serie^{l'})/CW_{i,i',l,l'} > \theta\}$
7: **end for**

Le programme calcule la matrice des contributions à partir de la formule des centres mobiles.

$$CW_{ii'}^{ll'} = \sum_{j=1}^{p} \frac{-W_{ii'}^{ll'}}{1 - W_{ii'}^{ll'}}(WX_{ij}^{l} + X_{i'j}^{l'})(2X_{ij}^{l} - \frac{2 - W_{ii'}^{ll'}}{1 - W_{ii'}^{ll'}}WX_{ij}^{l}) \tag{64}$$

Pour chaque instant (i, l) est calculé le vecteur $(WX_{ij}^{ll'})_{(j,l')}$. Nous ne refaisons pas pour chaque (i', l') le calcul de la variance suite à la mise à jour (suppression et renormalisation) de la matrice. Les arêtes étant sélectionnées, nous présentons ici la manière dont les matrices sont mises à jour pour l'apprentissage des appariements discriminants.

Mise à jour de la matrice de voisinage L'algorithme présenté ici gère la pénalisation des arêtes de la pioche et la répercussion de cette pénalisation. La distinction entre la version booléenne et la version progressive se fait à ce moment. Cet algorithme prend comme paramètre d'entrée la pioche, la matrice de voisinage ainsi que la matrice des contributions CW. Il renvoie une nouvelle matrice de voisinage mise à jour.

Approche booléenne Pour chaque instant de chaque série, l'algorithme cherche l'arête de la pioche qui maximise la contribution. Il s'assure que cette arête vérifie le critère de tolérance (ne rien faire si la pioche est vide) ; sinon, l'instant i est sorti de la boucle (il est mis dans l'ensemble i_{finies}). Il s'assure ensuite que l'arête vérifie le critère de neutralité (ne rien faire si toutes les arêtes appartiennent à la pioche), si ce n'est pas le cas, la série à laquelle appartenait l'arête est sortie de la boucle pour cet instant (elle est mise dans l'ensemble l'_{finies}).

Approche progressive Cette version prend en compte un argument supplémentaire, qui est la vélocité ν de l'algorithme. Pour chaque paire d'instants de la pioche, nous définissons un facteur de pénalisation fondé sur la contribution observée. Ce facteur est renormalisé par la contribution cumulée.

Il s'assure que cette arête vérifie le critère de tolérance (ne rien faire si la pioche est vide) ; si ce n'est pas le cas, l'instant i est sorti de la boucle (il est mis dans l'ensemble i_{finies}). Il n'y a pas de critère de neutralité dans le cadre progressif.

> **Remarque 59 : (Indépendance des instants)**
> *Dans les deux cas, les modifications se font en simultané sur toutes les lignes, en dépit du choix de l'approche locale. En effet, l'apprentissage du voisinage associé à chaque instant est indépendant des autres instants.*

Algorithm 5 mise-a-jour Version booléenne

Entrée : Pioche, classe cl, $CW_{(n_{cl}T\times n_{cl}T)}$,$W_{(nT\times nT)}$, lignes finies i_{fin}, blocs finis l'_{fin}
Sortie : matrice mise à jour $W_{(nT\times nT)}$, i_{fin},l'_{fin}
for $(l,i)\in[\![1,n_{cl}]\!]\times[\![1,T]\!]$ **do**
  if $i\notin i_{fin}$ **then**
    { critère de tolérance}
    if $\nexists(i,i',l,l')\in$PIOCHE **then**
      ajouter (i,l) à i_{fin}
    else
      Chercher i_1 et l_1 tels que $\begin{cases} CW_{i,i_1,l,l_1}\text{soit maximale} \\ (i,i_1,l,l_1)\in \text{PIOCHE} \end{cases}$
      { critère de neutralité}
      if $\forall i', (i,i',l,l_1)\in$ PIOCHE **then**
        ajouter (i,l,l_1) à l'_{fin}
      else
        $W_{ii'}^{ll'}=0$
      end if
    end if
  end if
end for

Algorithm 6 mise-a-jour Version progressive

Entrée : Pioche, classe cl, $CW_{(n_{cl}T\times n_{cl}T)}$,$W_{(nT\times nT)}$, lignes finies i_{fin}, vélocité ν
Sortie : $W_{(nT\times nT)}$, i_{fin}
for $(l,i)\in[\![1,n_{cl}]\!]\times[\![1,T]\!]$ **do**
  if $\nexists(i,i',l,l')\in$PIOCHE **then**
    ajouter (i,l) à i_{fin}
  else
    $(i,i',l,l_1)\in$ PIOCHE
    $W_{ii'}^{ll'}=(1-\nu\frac{CW_{ii'}^{ll'}}{\sum_{i_1,l_1}CW_{ii_1}^{ll_1}})W_{ii'}^{ll'}$
  end if
end for

Synthèse : algorithme LearnW pour l'apprentissage des appariements au sein des classes A présent que nous avons introduit toutes les étapes de l'algorithme, nous présentons ici la synthèse de l'algorithme adopté.

Algorithme final

Algorithm 7 Algorithme d'apprentissage intra LearnW

Entrée : Binit,X
Sortie : W
{Séparation de l'apprentissage classe par classe}
for $cl \in [\![1, nb - classe]\!]$ **do**
  W = initialisation[classe,Binit]
  {Définition du critère de tolérance}
  $\Theta = max_{i,i',l\in cl,l'\in cl} \|CONTRIB(X, W, i, i', serie^l, serie^{l'})\| \times \alpha$
  {Début de la boucle}
  repeat
    $(Pioche, C)$=select-aretes($W, X, classe, \Theta$)
    $(W, i_{fin}, {l'}_{fin})$= mise-a-jour$(Pioche, C, W, cl, i_{fin}, {l'}_{fin})$
  until $Pioche = \varnothing$
  W est normalisée en ligne
end for

La valeur prise pour α dépend naturellement du type d'algorithme. Dans la version booléenne, le terme α est très faible, de sorte que la condition sur α lors de la sélection des arêtes soit équivalente à une variation très faible de la variance.

Nous avons décrit l'algorithme d'apprentissage des appariements intra-classe. De manière symétrique, nous détaillons à présent les étapes de l'algorithme inter LearnB.

2.3 Apprentissage des appariements inter

Nous présentons dans cette section l'algorithme d'apprentissage des appariements inter. A nouveau, nous ferons la distinction lors de la mise à jour de la matrice, entre l'apprentissage progressif et l'apprentissage booléen. Du fait du choix d'un apprentissage entre paires de classes, la structure d'apprentissage est un peu modifiée.

Initialisation des matrices Cet algorithme d'initialisation prend comme paramètre d'entrée une structure de voisinage intra. Il renvoie en sortie la matrice B initiale. B est une structure de voisinage inter. Sa structure est assez proche de celle de l'algorithme intra-classe.

Algorithm 8 initialisation

Entrée : $Winit_{(nT\times nT)}$
Sortie : $B_{(nT\times nT)}$
for classe cl, séries $l, l' \in [\![1, K]\!]\times[\![1, n_{cl}]\!]^2$ **do**

$$B^{ll'} = \frac{1}{n - n_{cl}} \sum_{\substack{k \\ y_k \neq y_l}} Winit^{lk}$$

end for

Définition du critère de sélection des arêtes Cet algorithme de sélection des arêtes prend comme paramètre d'entrée une matrice d'appariement inter, ainsi que la matrice des valeurs $X = (X_1, .., X_p)$, où $X_k = (x^1_{1k}..x^1_{Tk}, x^2_{1k}.., x^n_{Tk})$. Il renvoie la pioche et la matrice des contributions C. Les contributions sont à nouveau calculées selon la formule des centres mobiles. Cette procédure parcourt l'ensemble des couples d'instants (i, i') de chaque paire de séries (l, l'), dans des séries de classes différentes.

Algorithm 9 select-aretes

Entrée : classes $cl1, cl2, B_{(nT\times nT)}$, $X_{(nT\times p)} = (X_1, .., X_p)$, θ
Sortie : Pioche, $C_{(n_{cl1}T\times n_{cl2}T)}$
Pioche = $\varnothing$
for $(l, i) \in [\![1, n_{cl1}]\!]\times[\![1, T]\!]$ **do**
  for $(l', i') \in [\![1, n_{cl2}]\!]\times[\![1, T]\!]$ **do**
    Calculer $C_{i,i',l,l'}$
    Ajouter à Pioche : $\{(i, i', l, l')/C_{i,i',l,l'} > \theta\}$
  end for
end for

Mise à jour de la matrice de voisinage Nous présentons dans ce paragraphe la mise à jour des poids dans le cadre de l'algorithme d'apprentissage inter-classes.

Approche booléenne Par symétrie avec l'algorithme intra, nous ne rentrons pas dans le détail des étapes de mise à jour qui sont similaires.

Algorithm 10 mise-a-jour version booléenne

Entrée : cl1,cl2, Pioche, $CB_{(n_{cl1}T\times n_{cl2}T)}$,$B_{(nT\times nT)}$, i_{fin}, l'_{fin}
Sortie : $B_{(nT\times nT)}$, i_{fin},l'_{fin}
for $(serie^l, i) \in [\![1, n_{cl1}]\!]\times[\![1, T]\!]$ **do**
  if $i \notin i_{fin}$ **then**
    if $\nexists (i, i', l, l') \in$PIOCHE **then**
      ajouter (i, l) à i_{fin}
    else
      Chercher i_1 et l_1 tels que $\begin{cases} C_{i,i_1,l,l_1} \text{soit maximale} \\ (i, i_1, l, l_1) \in \text{PIOCHE} \end{cases}$
      if $\forall i', (i, i', l, l_1) \in$ PIOCHE **then**
        ajouter (i, l, l_1) à l'_{fin}
      else
        $B^{ll'}_{ii'} = 0$
      end if
    end if
  end if
end for

Approche progressive

Synthèse : algorithme LearnB d'apprentissage des appariements entre les classes
Algorithme final

Algorithm 11 mise-a-jour version progressive

Entrée : Pioche, cl1,cl2 $CB_{(n_{cl1}T \times n_{cl2}T)}$,$B_{(nT \times nT)}$, i_{fin}, l'_{fin}
Sortie : $B_{(nT \times nT)}$, i_{finies},l'_{finies}
for $(l, i) \in [\![1, n_{cl1}]\!] \times [\![1, T]\!]$ **do**
  if $(i, i', l, l') \in$ PIOCHE **then**
    $B_{ii'}^{ll'} = (1 - \alpha \frac{CB_{ii'}^{ll'}}{\sum_{i_1, l_1} CB_{ii_1}^{ll_1}}) B_{ii'}^{ll'}$
  end if
end for

Algorithm 12 Algorithme d'apprentissage inter LearnB

Entrée : Winit,X
Sortie : B
{Séparation de l'apprentissage classe par classe}
for $cl1 \in [\![1, nb - classe]\!]$ **do**
  for $cl2 \in [\![1, nb - classe]\!]$ **do**
    B = initialisation[classe,Winit]
    {Définition du critère de tolérance}
    $\theta = max_{i,i',l \in cl1, l' \in cl2} \| CONTRIB(X, B, i, i', serie^l, serie^{l'}) \| \times \alpha$
    {Début de la boucle}
    repeat
      $(Pioche, CB)$=select-aretes($B, X, cl1, cl2, \theta$)
      (B, i_{fin}, l'_{fin})= mise-a-jour$(Pioche, CB, B, cl1, cl2, i_{fin}, l'_{fin})$
    until $Pioche = \varnothing$
    B est normalisée en ligne
  end for
end for

Après avoir détaillé les deux algorithmes intra et inter-classes, nous allons à présent expliciter la structure de l'algorithme fusionnant les deux approches, en vue de l'apprentissage d'appariements discriminants.

2.4 Apprentissage d'un bloc discriminant associé à la série

Suite à la définition de ces deux algorithmes, nous présentons maintenant une approche globale visant à apprendre simultanément des appariements intra et inter-classes, dans le but de produire, pour chaque série, un bloc discriminant. ***Algorithme final***

Algorithm 13 Algorithme d'apprentissage discriminant

Sortie : $(WB_1, ..., WB_n)$
{Initialisation}
W=W[J]
B=B[J]
repeat
  W = LearnW(W,X)
  B = LearnB(B,X)
until Stabilisation
$WB^{serie^l} = \frac{1}{n_{classe}} sum_{k/y_k = y_l} W^{lk}$

La stabilisation peut être par exemple "l'initialisation de B à l'étape t est la même que celle à l'étape t+1". En pratique, nous avons en général itéré la boucle repeat une seule

fois. L'apprentissage discriminant appris est l'ensemble des arêtes différentielles parmi celles qui sont les plus caractéristiques au sein d'un appariement complet. La stabilisation correspondante est donc "Nombre d'itération = k". Nous avons, dans cette section, présenté les algorithmes de manière détaillée. A partir de ces programmes, nous voulons à présent évaluer l'efficacité de la méthode proposée. Nous abordons donc dans la partie suivante la complexité des algorithmes introduits dans cette partie.

3 Complexité de l'algorithme

Nous allons dans cette section nous intéresser à la complexité des algorithmes. Les méthodes booléennes et progressives sont très différentes en ce qui concerne la finalité et la mise en application, et la complexité de l'algorithme d'apprentissage varie fortement entre les deux approches. C'est pourquoi nous distinguons les deux approches dans la suite.

3.1 Approche booléenne

L'approche booléenne conduit à peu d'arêtes concernées par chaque itération, mais les changements sont très forts et irréversibles. Ainsi, nous pouvons majorer le nombre d'itérations.

3.1.a Cas de l'algorithme intra

Rappelons la formule des centres mobiles utilisée pour calculer les contributions :

$$C(W)_{ii'}^{ll'} = \sum_{j=1}^{p} \frac{-W_{ii'}^{ll'}}{1-W_{ii'}^{ll'}}(WX_{ij}^{l} + X_{i'j}^{l'})(2X_{ij}^{l} - \frac{2-W_{ii'}^{ll'}}{1-W_{ii'}^{ll'}}WX_{ij}^{l}) \tag{65}$$

Les classes étant définies chacune par un bloc matriciel propre, nous pouvons les séparer au sein du processus. Ainsi, le calcul du produit matriciel WX est alors d'une complexité en $O(p\sum_k I_k(n_kT)^2)$. C'est le terme dominant dans le calcul de la matrice des contributions, l'algorithme inter se fait donc en $O(p\sum_k I_k(n_kT)^2)$, où I_k est le nombre d'itérations impactant la classe k. Dans le pire des cas, $I_k = (n_k-1)(T-1)$. Donc, dans le pire des cas, la complexité de l'algorithme intra est en $O(pT^3\sum_k (n_k)^3)$

3.1.b Cas de l'algorithme inter

Les paires de classes étant définies chacune à nouveau par un bloc matriciel propre, nous pouvons les séparer au sein du processus. Ainsi, le calcul du produit matriciel WX est alors d'une complexité en $O(p\sum_{k_1,k_2,k_1\neq k_2} I_{k_1,k_2}n_{k_1}n_{k_2}T^2)$. C'est le terme dominant dans le calcul de la matrice des contributions, l'algorithme inter se fait donc en $O(p\sum_k I_k(n_kT)^2)$, où I_{k_1,k_2} est le nombre d'itérations impactant la classe k_1 par rapport à la classe k_2. Dans le pire des cas, $I_{k_1,k_2} = (n_{k_2}-1)(T-1)$. Donc, dans le pire des cas, la complexité de l'algorithme intra est en $O(pT^3\sum_{k_2,k_1\neq k_2}(n-n_{k_2})n_{k_2}{}^2)$ $=O(pnT^3\sum_k n_k^2)$.

Notons que la complexité de l'algorithme considérant les apprentissages inter entre une classe et toutes les autres classes simultanément a une complexité bien plus forte en $O(pT^3n^3)$

3.1.c Cas de l'algorithme global

La complexité de l'algorithme global est délicate à calculer. Nous avons vu que le processus se stabilise en au plus T itérations ; cependant, ce cas correspond à une moindre pénalisation d'arêtes, donc, au cas le plus favorable de l'algorithme intra (au plus n itérations). La complexité de l'algorithme global est alors en $O(n^3T^3)$. Ainsi, en théorie, la complexité de l'algorithme global est donc majorée par $O(T^4n\sum_k n_k^2)$. En pratique, la pénalisation de n arêtes entraine nécessairement la disparition d'au moins une arête dans le bloc moyen. La complexité générale de l'algorithme reste bien en $O(T^3n\sum_k n_k^2)$.

Nous voyons que la complexité de l'algorithme d'apprentissage booléen des matrices d'appariement a une complexité équivalente aux méthodes d'apprentissage classiques pour de tels problèmes d'optimisation, à l'instar de la méthode du gradient projeté par exemple. L'approche progressive n'est pas équivalente, a priori, car le nombre d'itération n'est pas borné. Nous calculons dans la suite la complexité de l'approche progressive.

3.2 Approche progressive

Cette seconde méthode qui consiste à faire des petites modifications des poids au sein de la matrice d'appariement conduit à des poids jamais nuls. Ainsi, l'algorithme converge lorsque les modifications sont faibles. A priori, on ne peut pas borner ce nombre de modification. Cependant, nous allons voir que ce fait n'a que peu d'incidence sur la complexité de l'algorithme.

3.2.a Cas de l'algorithme intra

Dans le cadre de l'approche progressive, le calcul de la matrice des contributions est le même, et l'algorithme intra se fait en $O(p\sum_k I_k(n_kT)^2)$ où I_k est le nombre d'itérations impactant la classe k. Ici, le nombre d'itération n'est plus borné. Cependant, du fait du caractère géométrique de la décroissance, la convergence est assez rapide. La complexité de l'algorithme intra est en $O(pT^2\sum_k (n_k)^2)$.

3.2.b Cas de l'algorithme inter

De la même façon, le calcul de la matrice des contribution se fait en $O(p\sum_k(n_kT)^2)$. Dans le pire des cas, $I_{k_1,k_2} = (n_{k_2} - 1)(T - 1)$. Donc, dans le pire des cas, la complexité de l'algorithme intra est en $O(pT^3\sum_{k_2,k_1\neq k_2}(n - n_{k_2})n_{k_2}{}^2)$ $=O(pnT^3\sum_k n_k^2)$.

3.2.c Cas de l'algorithme global

Dans ce cas, l'algorithme global n'est plus sûr de converger. En pratique, l'algorithme ne tournera que pour un nombre très faible d'itérations. Sa complexité sera de l'ordre de celle de l'algorithme inter.

La complexité de l'algorithme d'apprentissage progressif des matrices d'appariement a une complexité équivalente à celle du précédent. En effet, si le nombre de modifications n'est plus borné a priori, du fait du caractère géométrique des pénalisations, la convergence est rapide et le nombre d'itération relativement faible.

Nous avons vu que la méthode que nous proposions donnait des résultats comparables aux méthodes de l'état de l'art en termes de complexité. Nous allons à présent présenter à partir

des jeux de données étudiés précédemment les résultats des algorithmes d'apprentissage. Nous étudions dans la suite l'algorithme booléen, dont nous avons vu que la complexité était équivalente à celle de l'algorithme progressif, et qui présente l'avantage de ne pas avoir un paramètre à étudier, à savoir l'intensité des pénalisations.

4 Etude des appariements appris

Nous allons donc dans cette section étudier sur un exemple l'algorithme présenté ci-dessus. Nous nous intéresserons dans un premier temps aux matrices d'appariement apprises à l'issue du processus d'apprentissage, et nous étudierons la stabilité des algorithmes pour des petites perturbations. Nous nous limitons dans cette section, au cas de l'algorithme d'apprentissage booléen. Les jeux de données considérés sont ceux présentés dans l'annexe C. Ils présentent l'intérêt de proposer un découpage en classes définies par une signature particulière, en permettant de contrôler la structure des événements discriminants. En particulier, les séries aux seins des classes présentent des profils très différents, et présentent des similarités entre les classes.

4.1 Allure des appariements

Les appariements intra appris ont permis de supprimer les arêtes ayant une forte variabilité au sein d'une classe. Ainsi, les zones qui demeurent actives au sein des appariements initiaux du processus inter sont celles qui appartiennent à des régions relativement uniformes au sein de la classe. Les appariements inter appris ont permis de supprimer les arêtes faisant état d'une faible variabilité au sein d'une classe. De même, les zones qui demeurent actives au sein des appariements initiaux du nouveau processus intra sont celles qui appartiennent à des régions différentiant fortement les différentes classes.

La figure 25 visualise les poids des blocs appris au sein de la classe Begin à partir du couplage complet initial. Nous avons sélectionné quatre séries de la classe Begin, dont nous comparons les appariements appris, deux ont leur bosse centrale positive, les deux autres l'ont négative. La figure visualise en clair les zones couplées avec un poids fort, et en foncé, les zones couplées avec un poids faible ou nul. Nous remarquons le fait que dans chaque configuration, apparaît le couplage des petites cloches initiales (petit carré clair), qui caractérise la classe Begin. Le plateau central est lié dans certains cas. Le plateau final, important pour discriminer la série avec les séries des autres classes n'est jamais présent. Ce point illustre l'importance de prendre en considération les événements discriminants.

De manière symétrique, pour les blocs de la classe End, les comportements sont similaires.

Enfin, pour la classe Middle, nous obtenons une structure en damier. En effet, la structure minimisant la variabilité consiste à se limiter aux instants dont les valeurs sont proches.

Cependant, dans les trois cas, les couples d'instants qui ressortent avec un poids important, s'ils sont caractéristiques des classes, ne sont pas discriminants. Observons alors l'impact de l'initialisation par les blocs inter appris. Nous présentons alors, sur les figures 28 et 29 pour un bloc représentant deux séries de la classe Begin, l'évolution des blocs appris au fil des itérations. Chaque ligne correspond à une itération. Sur la première colonne, nous observons la structure initiale, et en dernière colonne, la structure à l'issue de l'apprentissage. Nous remarquons que dès la seconde itération, le processus est stabilisé. Il fait ressortir le

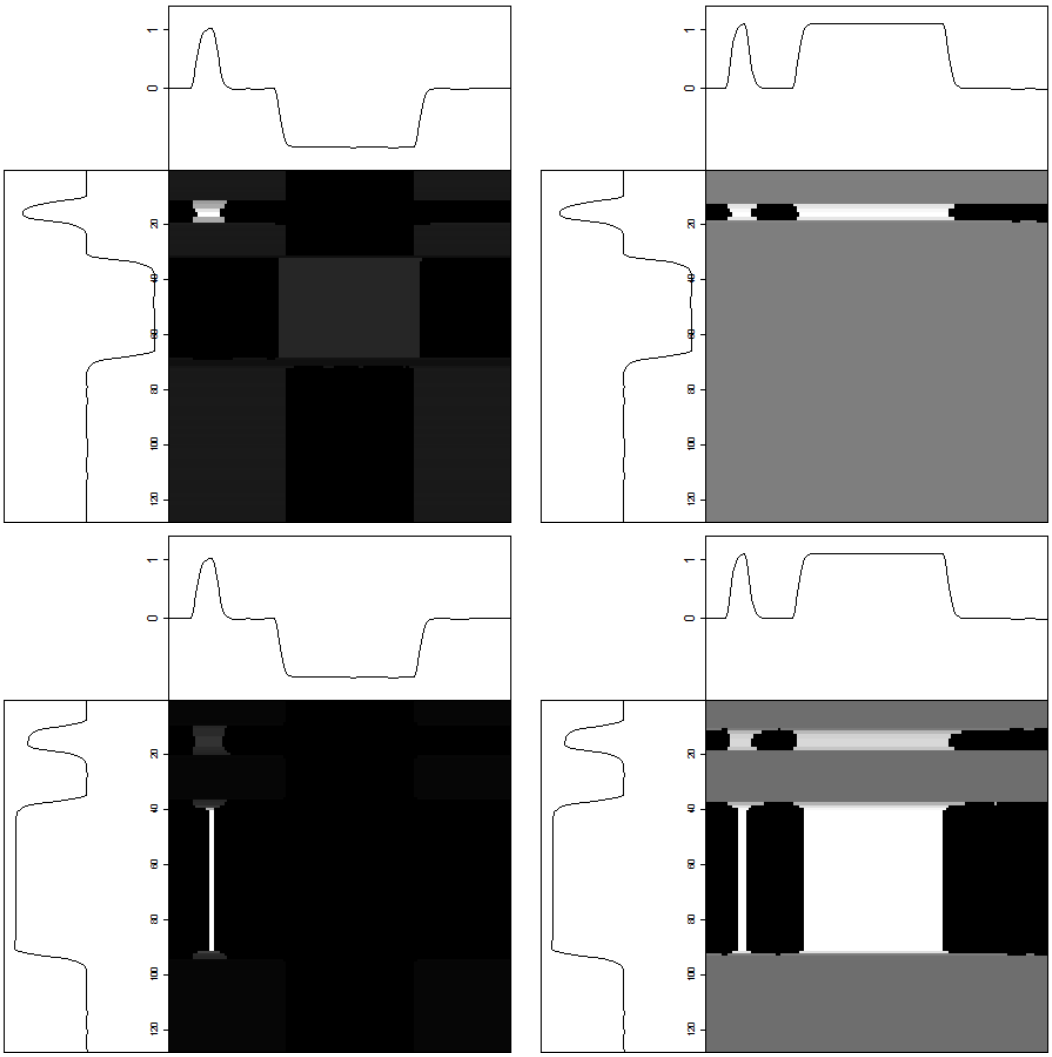


FIGURE 25 – Blocs intra appris pour la classe Begin

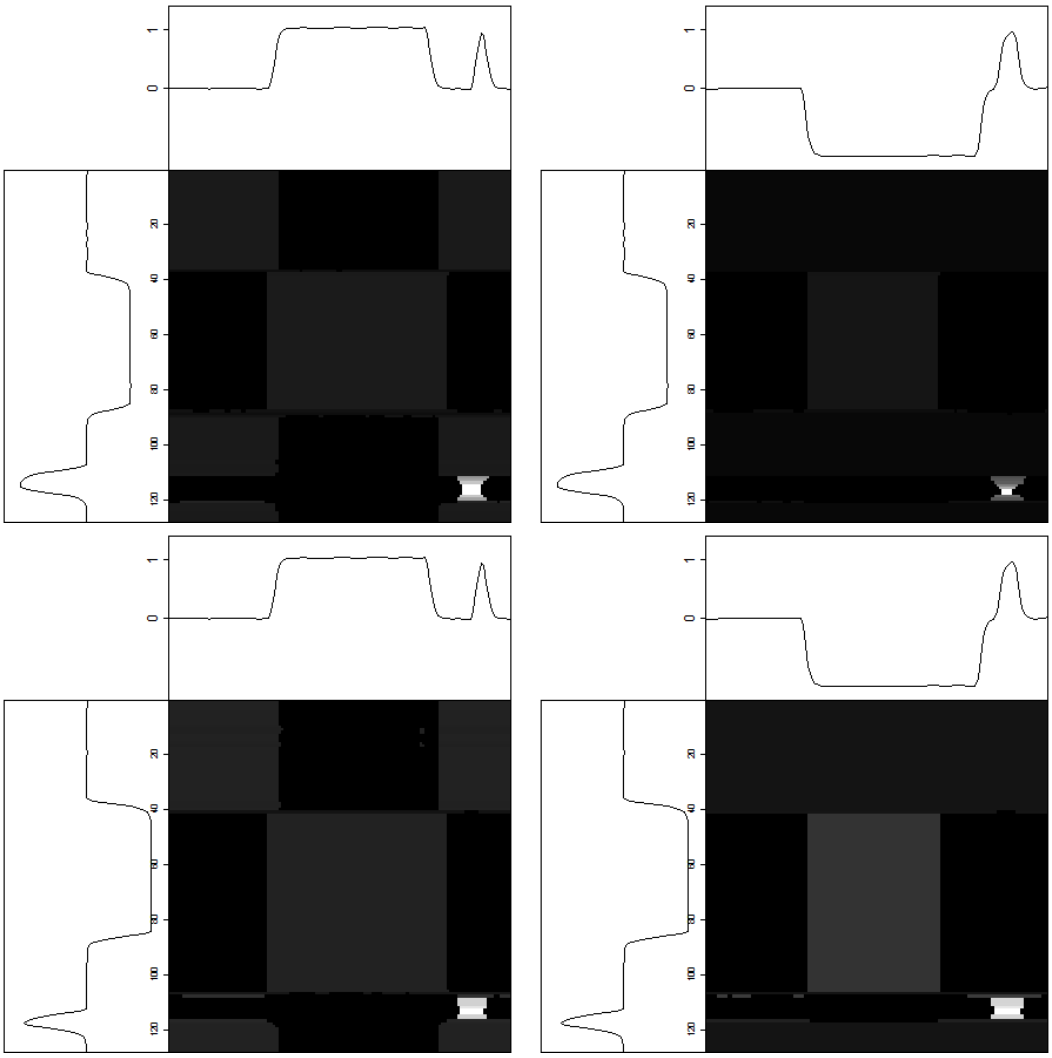


FIGURE 26 – Blocs intra appris pour la classe End

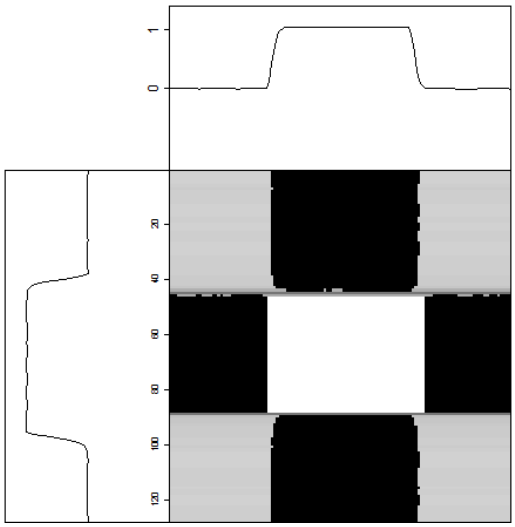


FIGURE 27 – Bloc intra appris pour la classe Middle

couplage des petites bosses initiales, mais également la zone du second plateau, couplée aux instants inter-plateaux, qui est une zone importante pour la discrimination. La combinaison des approches intra et inter a ainsi permis un apprentissage discriminant.

Nous avons observé que l'allure des blocs discriminants était préservée après un très faible nombre d'itérations de l'algorithme global. Les résultats semblent converger vers une structure correspondant aux attentes. Nous allons cependant vérifier que le processus d'apprentissage parvient à apprendre des matrices d'appariement ayant une variabilité intra minimisée et une variabilité inter maximisée. Pour cela, nous observons l'évolution de ces deux variances au cours des itérations. Nous nous assurerons également de la convergence de l'algorithme discriminant global lors de la fusion des deux approches.

4.2 Stabilité et convergence de l'algorithme

Nous allons voir dans cette sous-partie que l'approche choisie assure naturellement la convergence des algorithmes LearnW et LearnB vers une structure de voisinage au pouvoir discriminant augmenté.

4.2.a Diminution de la variance intra et augmentation de la variance inter

Dans le cadre intra, le processus itératif assure à chaque étape une diminution de la variance. De par la définition de la contribution, une arête appartenant à la pioche ayant une contribution positive, sa suppression entraîne nécessairement une diminution de la variance. Cette diminution est assurée par le choix de la pénalisation définitive, mettant à 0 le poids d'une unique arête de la pioche, choisie pour être pénalisée. Cette approche évite les effets de bord dus d'une part à la pénalisation simultanée de plusieurs arêtes, et d'autre part, ceux dus au cas rare de changements de signes évoqués au paragraphe 3.3.a du chapitre 3. Ainsi, nous sommes assurés d'avoir au fil des itérations du processus LearnW une variance intra plus faible que la variance intra initiale. De manière symétrique, nous sommes assurés d'avoir au fil des itérations du processus LearnB une variance inter plus grande que la variance inter initiale. Le processus global n'assure a priori pas l'augmentation du pouvoir discriminant. En effet, lors de l'initialisation, le fait de faire une moyenne et un seuillage n'assure plus la décroissance (respectivement la croissance) de la variance intra (respectivement inter). Nous pouvons toutefois ajouter cela comme une condition d'arrêt dans l'algorithme discriminant. Si, à la fin d'un des processus intra ou inter, le pouvoir discriminant est plus faible qu'à la fin du processus précédent, le programme s'arrête et renvoie la matrice précédente.

Nous observons le phénomène de décroissance et de stabilisation en fin de processus. La diminution de la pente que nous observons autour de la 500^e itération traduit le fait que l'apprentissage s'est stabilisé pour certaines classes.

Considérons les différentes itérations des deux processus LearnW et LearnB. Considérons dans un premier temps le processus intra. Lors des deux itérations suivantes du processus global, la variance intra initiale est plus grande que la variance intra obtenue à la fin du processus de l'itération précédente. Ceci est dû au fait que l'initialisation par des blocs sémantiques apporte de la variabilité. Des arêtes supprimées au préalable sont réinitialisées avec un poids positif. La chute de variance intra se poursuit, avec une convergence plus rapide, du fait d'une initialisation avec une matrice initiale plus creuse. Nous remarquons que la valeur de la variance trouvée est plus faible. L'enchaînement en cascade des deux processus entraîne donc

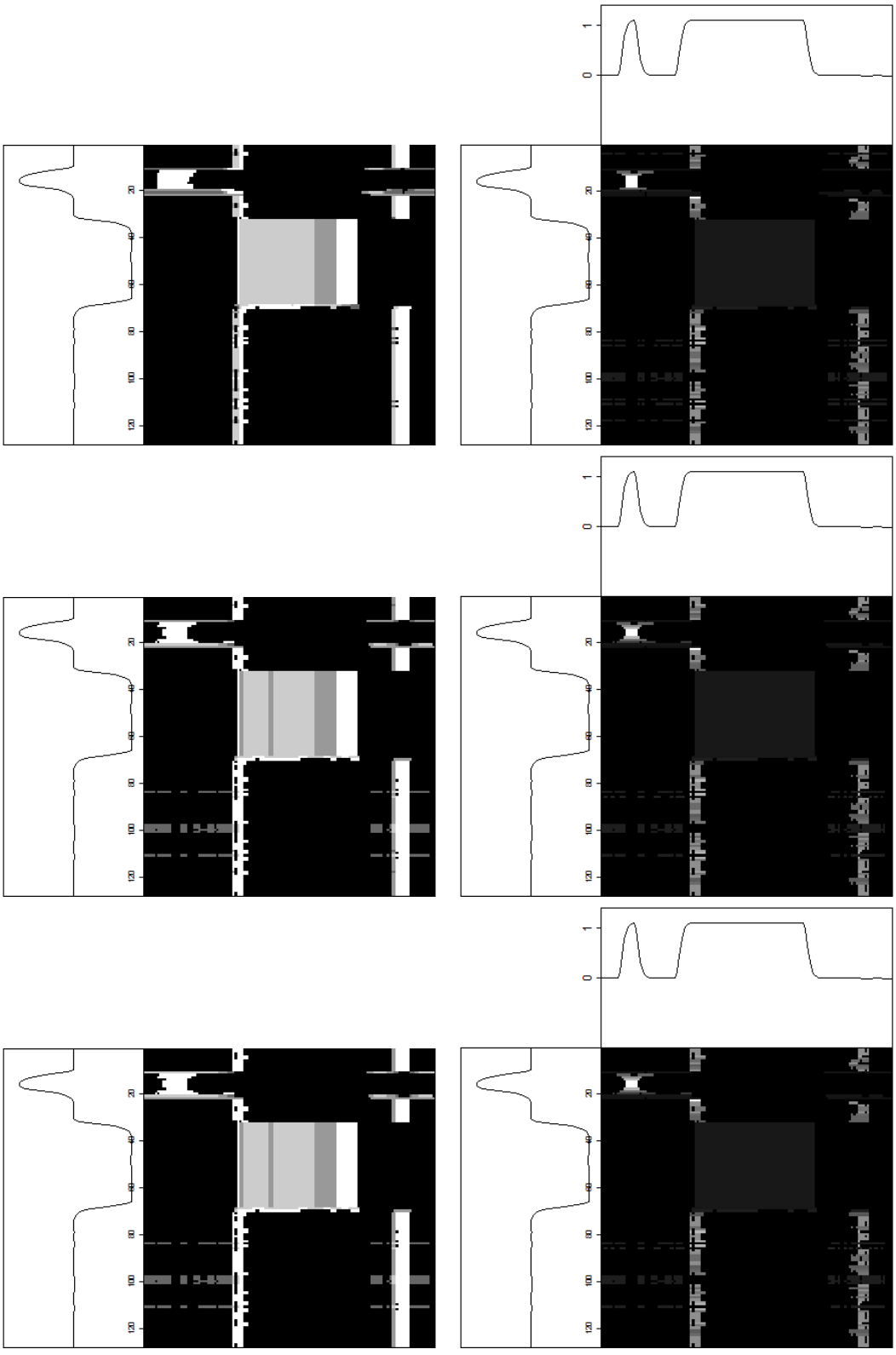


FIGURE 28 – Evolution des blocs intra : première itération

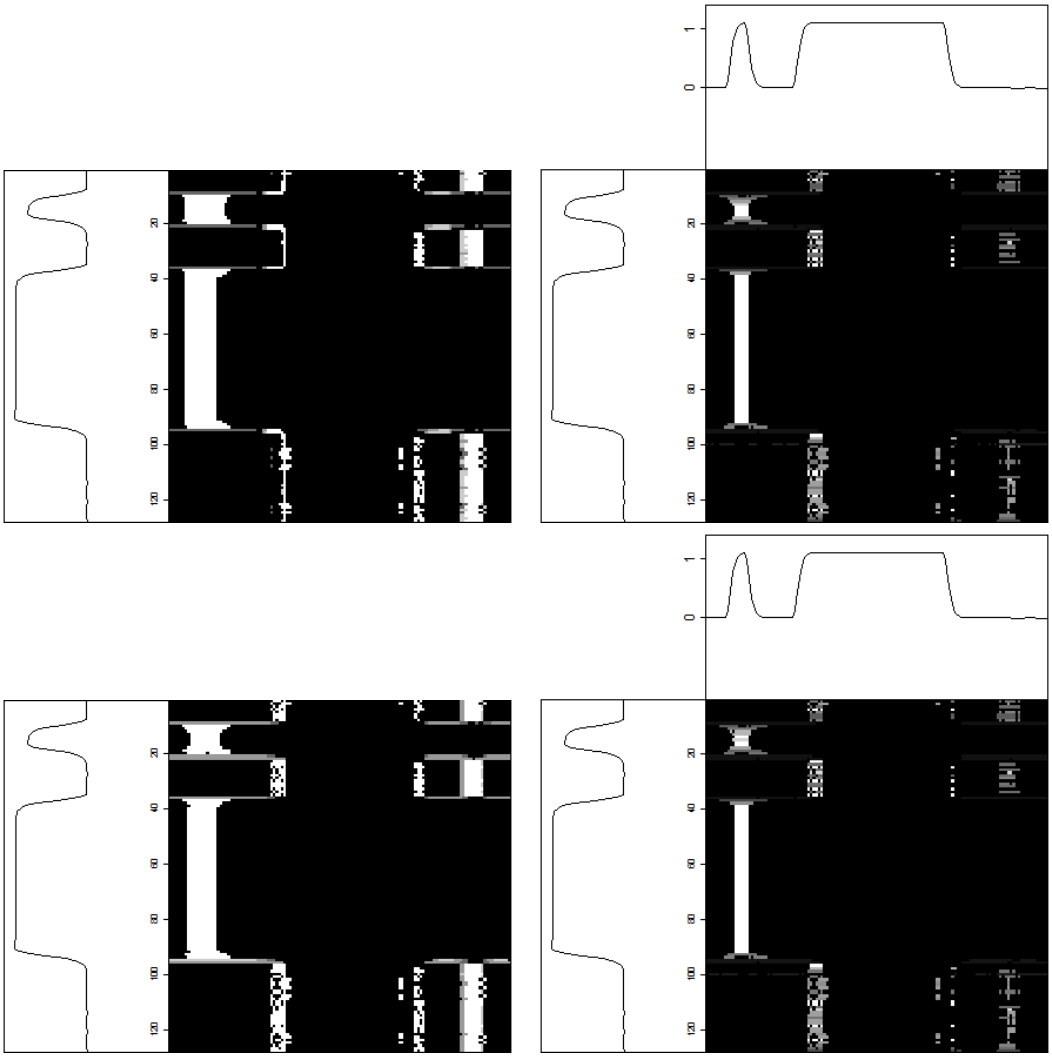


FIGURE 29 – Evolution des blocs intra : seconde itération

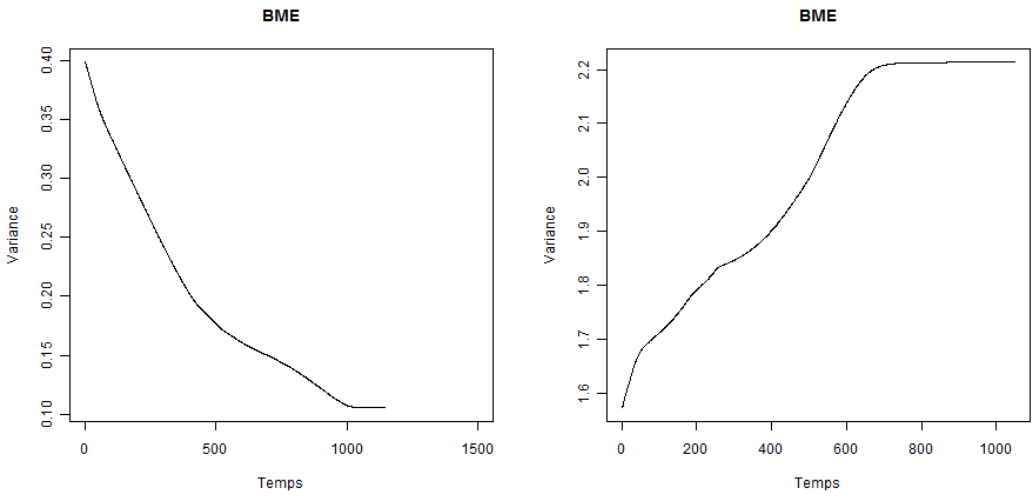


FIGURE 30 – Croissance et décroissance à la première itération de LearnW et LearnB

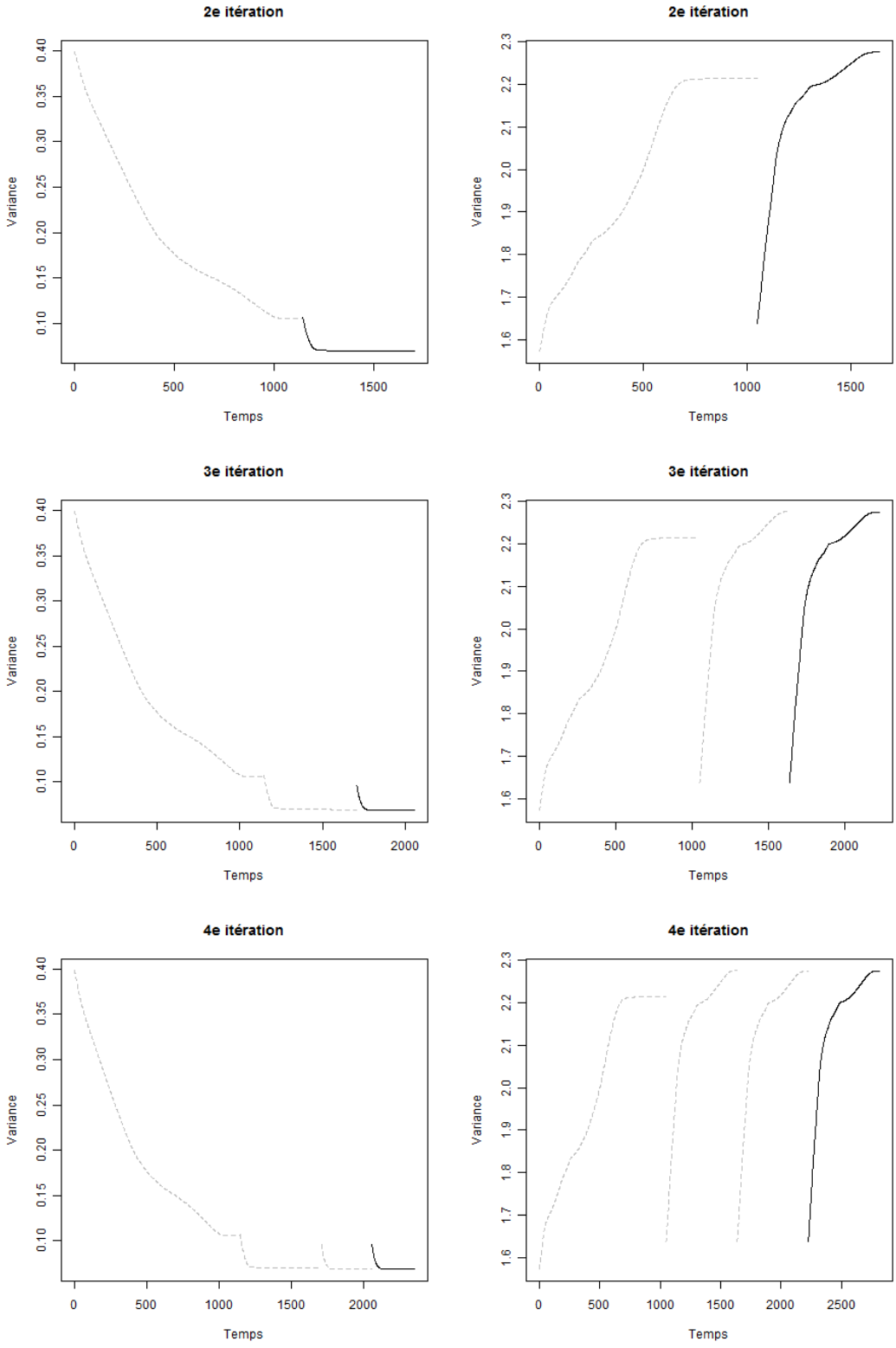


FIGURE 31 – Croissance et décroissance aux deux itérations suivantes de LearnW et LearnB

un affinement des appariements appris. De manière symétrique, la variance inter augmente suite à l'enchaînement des processus. La très grande similitude des deux courbes correspondant aux variances inter pour les deuxième et troisième itérations montre la convergence de la structure apprise. Pour le jeu BME, nous constatons notamment qu'à l'issue de 3 itérations, la variance a convergé.

4.2.b Arrêt du processus itératif

La convergence des algorithmes d'apprentissage LearnW et LearnB est assurée par le fait de pénaliser totalement une arête de la pioche à chaque itération. Ainsi, une arête de la pioche sera supprimée à chaque itération jusqu'au vidage intégral de la pioche. Aucune arête supprimée n'a de risque de revenir plus tard au cours du processus d'apprentissage. L'algorithme LearnW s'arrêtera donc avant moins de $Tmax_k(n_k)$ itérations, une arête au plus étant supprimée sur chaque ligne à chaque itération. Par symétrie, il en est de même pour l'algorithme inter LearnB. Il s'arrêtera avec au plus $Tmax_k(n - n_k)$ itérations.

Le processus combinant les deux approches converge également vers une matrice constante. D'une itération à l'autre, des arêtes ne peuvent qu'être supprimées. En effet, une arête qui a été supprimée à un moment ne peut pas réapparaître. Ainsi, si aucune arête n'est supprimée au cours d'une itération, le processus renvoie à l'initialisation la matrice initiale précédente. Ceci nous assure de toujours converger. Cependant, rien ne nous assure du caractère optimal des appariements finaux appris lors du processus croisé. Il est parfois préférable de s'arrêter en cours d'algorithme. Ceci justifie l'ajout de la condition d'arrêt spécifique.

Conclusion

Le chapitre 3 avait introduit deux algorithmes pour l'apprentissage d'appariements intra et inter-classes. Le présent chapitre en a fait une étude plus détaillée, et a explicité une méthode pour coupler les deux algorithmes, en vue de la recherche d'un appariement discriminant. Nous avons observé que l'approche proposée est de complexité équivalente aux algorithmes usuels pour de tels problèmes d'optimisation (tel l'algorithme du gradient projeté). Nous avons également observé sur des données simulées l'allure des appariements appris. La combinaison des deux approches permet de faire ressortir des éléments discriminants du jeu de données, tels que souhaités. Il faudrait à présent tester les méthodes d'apprentissage sur des problématiques de classification, en se penchant sur des problématiques réelles. Nous introduirons donc dans la partie suivante des jeux réels, et confronterons les appariements appris à des tâches de classification.

Conclusion de la partie II

Nous avons proposé, dans cette partie, une méthode pour l'apprentissage d'appariements discriminants. L'apprentissage des appariements se formalise comme un problème d'optimisation. L'approche adoptée pour résoudre ce problème consiste à renforcer et à pénaliser les arêtes de manière itérative, en fonction de leur contribution à la variabilité intra et inter-classes. Nous avons montré que les méthodes de gradient usuellement utilisées sont équivalentes à notre approche en termes de complexité. Nous avons également étudié la stabilité des algorithmes et étudié sur des jeux simulés les résultats de ces approches. Nous allons voir dans la partie suivante l'efficacité des métriques fondées sur ces appariements pour la classification de séries temporelles.

Partie III

Utilisation des appariements appris

Au regard de l'importance de l'apprentissage d'appariements adaptés à la discrimination d'une partition de séries temporelles pour de nombreuses tâches, à l'instar de la classification de séries temporelles, nous avons introduit un algorithme d'apprentissage d'appariements discriminants fondé sur la maximisation de la variance inter et la minimisation de la variance intra.
Nous proposons dans cette partie de définir à partir des appariements appris, une nouvelle métrique pour la classification, et de montrer son efficacité dans le cadre d'une classification k-NN de données de consommation électrique. Le premier chapitre consiste à définir la métrique. Cette métrique utilise en particulier l'entropie de Shannon pour la pondération des instants. Dans le second chapitre, nous observons les résultats induits par cette métrique dans le cadre de la classification par k plus proches voisins de séries issues d'un jeu de données réelles de consommation électrique. Nous comparons nos résultats à ceux des approches fondées sur les métriques classiques.

Chapitre 5

Stratégies de pondération des instants de la série en fonction de leur caractère discriminant

Nous souhaitons dans ce chapitre appliquer les appariements appris à la discrimination et la classification de séries temporelles. Nous définissons tout d'abord dans ce chapitre un premier système de pondération des instants en fonction de leur caractère discriminant. Ces poids sont définis par rapport à la variabilité du voisinage de l'instant. Nous définissons dans un second temps un autre système de poids fondé sur la notion d'inertie de Shannon. A partir de ces deux systèmes de poids, nous proposons une nouvelle métrique discriminante en vue de la classification de séries.

Nous avons vu dans le chapitre 2 de la partie I que la définition de la variance repose sur le choix d'une métrique, dont nous avons proposé une méthode d'apprentissage au chapitre 3 de la partie II. Nous avons alors observé que les appariements obtenus étaient discriminants, au sens d'une augmentation de la variance inter et d'une diminution de la variance intra. Nous souhaitons, à partir d'une matrice d'appariement donnée, classique (Euclidienne ou DTW par exemple) ou apprise par le processus décrit au chapitre 4 de la partie II, définir une métrique qui extraie une information sur l'aspect discriminant d'un instant. Dans le cadre de l'appariement euclidien ou de l'apprentissage booléen, par exemple, un même poids est donné à chaque instant ; l'objectif de l'apprentissage est de rechercher les arêtes les plus discriminantes associées à chaque instant. Cependant, certains instants sont plus aptes à discriminer les séries. L'objectif de cette partie est de définir un système de poids sur les instants des séries en fonction de leur caractère discriminant. Les poids que nous proposons d'apprendre sont définis à partir d'une matrice d'appariement quelconque. Nous étudierons les poids pour une matrice quelconque dans un premier temps, puis nous appliquerons à la fin ces poids aux matrices d'appariement apprises.

Tout au long de ce chapitre, les résultats sont illustrés par des séries issues des deux jeux de données BME et UMD, détaillés en annexe, et dont nous rappelons dans les figures 32 et 33 l'allure des courbes.

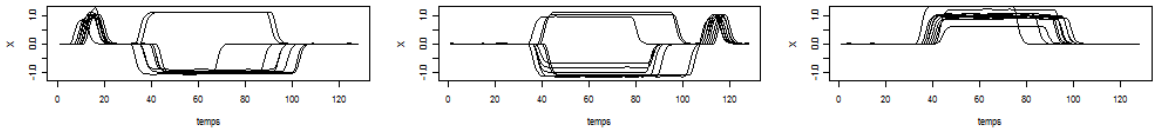


FIGURE 32 – Jeu BME

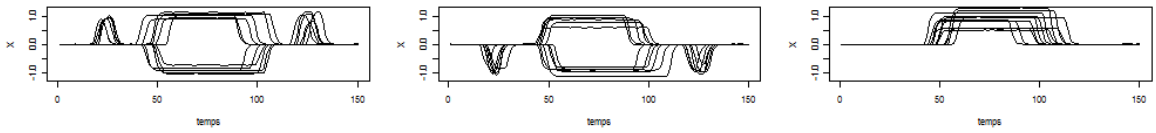


FIGURE 33 – Jeu UMD

1 Utilisation de la variance en vue de la définition d'instants discriminants

Les matrices d'appariement définissent un système de poids associé à chaque instant. Ainsi, tous les instants ont même poids dans l'apprentissage. Or, certains instants sont plus discriminants. Par exemple, la petite bosse des jeux UMD et BME est plus discriminante que la grande. Nous proposons plusieurs approches pour définir des poids discriminants. La première approche est fondée sur la variance des instants. Nous considérons que les instants les plus discriminants sont ceux les plus centraux dans la classe, c'est-à-dire les instants induisant le moins de variabilité au sein de la classe et le plus de variabilité entre les classes.

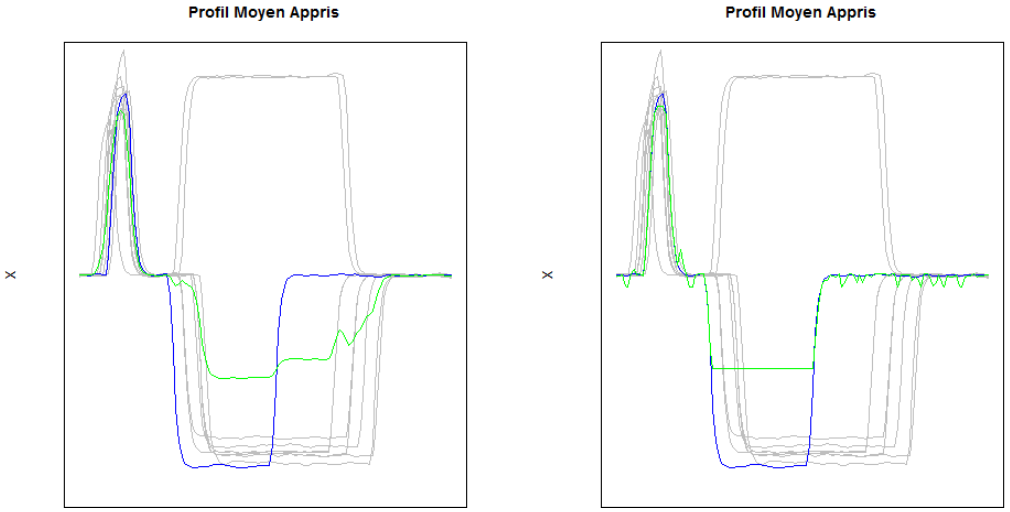
1.1 Notion de profil moyen

Nous avons vu, à travers la formule des centres mobiles, que la variance associée à un voisinage se définissait comme moyenne des carrés des écarts à la moyenne des valeurs prises dans ce voisinage. Ainsi, à chaque instant de chaque série est associé la moyenne des valeurs prises dans le voisinage de cet instant. Nous appellons profil moyen associé à la série S_l et au voisinage M cette série, et nous le notons $\mathbb{S}^{M^l} = M^l_{i\cdot}X$.

Le profil moyen associé au voisinage intra $\mathbb{S}^{W^l}$ est une série approchant l'allure de S^l en fonction des éléments caractéristiques de la classe, et le profil moyen associé au voisinage inter $\mathbb{S}^{B^l}$ est une série éloignée au maximum de la série S^l selon les éléments les plus différentiels.

D'une manière générale, si un événement est commun à chaque série de la classe, on trouve un voisinage pour lequel la valeur du profil moyen associé est proche de la valeur de la série. Si un événement ne trouve pas écho au sein des autres séries de la classe, il s'éloigne fortement de son profil moyen.

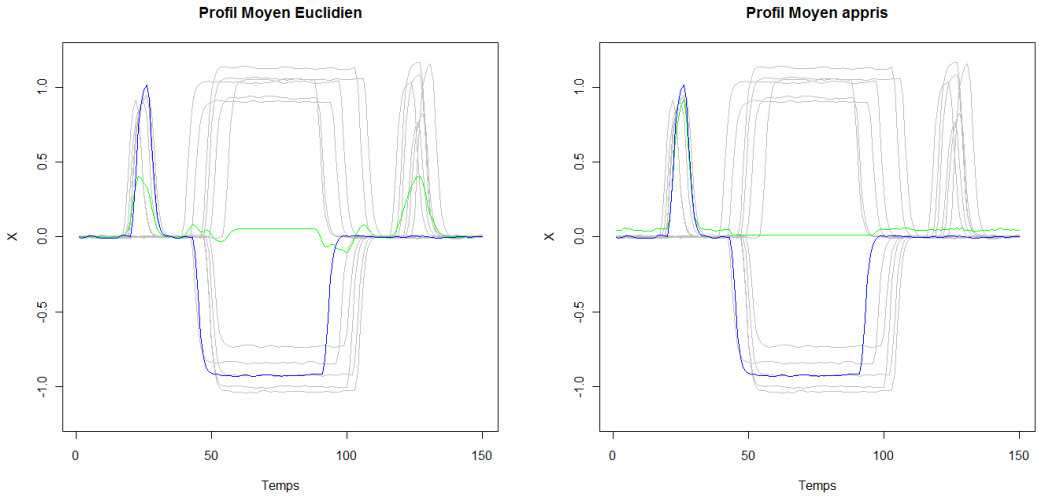
La figure 34 donne pour le jeu BME présenté en annexe, le profil moyen tantôt associé à un couplage euclidien, tantôt associé à l'appariement intra-classe appris par notre méthode. Notons que le profil moyen euclidien est commun à toutes les séries de la classe et correspond à une moyenne usuelle instant par instant de toutes les séries. L'appariement appris est propre à chaque série. Il cherche à approcher la série initiale, en tenant compte de la structure de la classe.



(a) Appariement euclidien (b) Appariement appris

FIGURE 34 – Profil moyen de la classe Begin du jeu BME

Dans le contexte du jeu BME, la petite bosse initiale a été reconnue comme un élément caractéristique de la classe pour cette série, tant par l'alignement euclidien que par l'appariement appris par notre approche. La bosse centrale n'est pas très bien approchée par le profil moyen, du fait qu'elle apparaît tantôt vers le haut, tantôt vers le bas, et n'est en cela pas caractéristique. Dans le cas de l'approche apprise, la structure de la courbe est plus proche de celle du profil moyen.



(a) Appariement euclidien (b) Appariement appris

FIGURE 35 – Profil moyen de la classe Up du jeu UMD

La figure 35 donne le profil moyen pour le jeu UMD présenté en annexe. Dans le contexte du jeu UMD, l'apprentissage a reconnu la petite bosse initiale comme un élément caractéristique de la classe pour cette série, tandis que l'appparieemment euclidien lui donne une hauteur très faible. La bosse centrale n'est pas reconnaissable, du fait qu'elle apparaisse tantôt vers

le haut, tantôt vers le bas et n'est en cela pas caractéristique.

Le profil moyen est une moyenne pondérée des observations, et correspond à un indicateur de position. A l'instar de la moyenne, nous souhaiterions utiliser les profils moyens pour la définition d'un indicateur de dispersion. Le profil moyen joue un rôle important car il permet de redéfinir la variance associée à une structure de voisinage de la manière suivante :

$$V_W = \frac{1}{nT} \sum_{\substack{i\in\{1..T\} \\ l\in\{1..n\}}} (S^l - \mathbb{S}^{W\,l})^2$$

$$V_B = \frac{1}{nT} \sum_{\substack{i\in\{1..T\} \\ l\in\{1..n\}}} (S^l - \mathbb{S}^{B\,l})^2$$

La variance intra (respectivement inter) est la moyenne des valeurs prises pour chaque instant de chaque série par les écarts au carré entre la série et son profil moyen intra (resp inter) associé au sein de la classe, i.e., la moyenne des écarts au carré pour chaque instant de chaque série, entre la valeur prise en cet instant et la moyenne (pondérée dans le cas d'une approche progressive, non pondérée dans le cas booléen) des valeurs prises par tous les instants voisins de celui-ci.
Nous souhaitons minimiser la variance entre les séries de chaque classe et maximiser la variance entre les séries de classes différentes.

1.2 Variabilité d'un instant

Le travail effectué sur la généralisation de la variance, présenté au chapitre 2 de la partie I a permis de généraliser la notion de variance intra et de variance inter à un ensemble de données liées à la fois par une structure de contiguïté et par une structure de découpage en classe. Rappelons sommairement la formule de décomposition usuelle de la variance dans le cadre d'une structure d'individus regroupés en classes. La variance usuelle (variance dite totale) est la somme de la variance intra-classe et de la variance inter-classes. La variance intra-classe usuelle est la moyenne des écarts au carré entre les individus et le centre de la classe. La variance inter usuelle est égale à la variance des moyennes des classes.

Si nous considérons une structure de voisinage intra et une structure de voisinage inter complémentaires, les variances associées à ces deux structures de voisinage, telles qu'elles ont été redéfinies au chapitre 2 de la partie I, représentent une variance dite totale. De ce fait, l'appellation de variance inter prend sens ici.
Cependant, de manière générale, le profil moyen est une moyenne des observations de la classe. La variance associée à une structure de voisinage est donc une généralisation de la notion de variance intra, qu'elle soit associée à une structure de voisinage intra ou inter. En effet, il s'agit de la moyenne des écarts des observations au profil moyen.
La variance associée à la structure de voisinage inter peut ainsi être vue tant comme une généralisation de la variance intra que de la variance inter usuelles. Cette dualité associée à la variance inter est très importante, car de ce fait, les structures de voisinage sont symétriques.

Cependant, dans le calcul de la variance usuelle, la moyenne est commune à tous les objets de la classe et construite à partir de tous les instants équi-pondérés. Dans notre approche, nous définissons une moyenne mobile pour chaque individu, à partir d'un système de poids pouvant privilégier certaines arêtes. Le profil moyen correspond à un centrage local vis-à-vis

du voisinage de l'instant. La valeur du profil moyen est une moyenne pondérée des valeurs prises par les autres instants. A partir d'une structure de voisinage, nous pouvons alors définir la variabilité du voisinage de chaque instant. A l'instar de la variance d'un ensemble, qui consiste à calculer globalement l'écart à la moyenne de voisinage, nous nous penchons localement sur la variance d'un instant, qui est la moyenne des écarts au carré pondérés entre ses instants voisins et la moyenne de voisinage, le pendant de la variance intra dans le cadre d'une structure de voisinage où le voisinage d'un instant constitue une classe (recouvrante).

Dans ce qui suit, les voisinages sont en général constitués de chacune des autres séries. Nous verrons dans la partie suivante comment nous pouvons généraliser cette variabilité à un sous-ensemble de séries.

1.2.a Variabilité du voisinage associé à un instant

Nous nous intéressons ici à la définition d'une variance associée à une structure de voisinage quelconque. La variabilité d'une série S_l au sein d'un voisinage M, qu'on note ${\sigma^M}_l$ se définit comme

$$ {\sigma^M}_l[i] = \sum_{l' M^{ll'} \neq 0} \sum_{j=1}^{l} tM_{ij}^{ll'} \left(S_j^{l'} - \bar{S}_{li}^{\,M} \right)^2 \tag{66} $$

où ${\mathbb{S}^l}^M$ est le profil moyen de S_l calculé sur la base du voisinage M.

$$ {\mathbb{S}^l}^M[i] = \sum_{l' M^{ll'} \neq 0} \sum_{j=1}^{l} tM_{ij}^{ll'} S_j^{l'} \tag{67} $$

On nomme profil variance la série temporelle ${\sigma^M}_l$. Cette série s'apparente à un terme de variance intra dans la décomposition usuelle de la variance. Le profil variance ${\sigma^M}_l$ prend des valeurs faibles lorsque les arêtes permettant de définir le profil moyen s'éloignent peu de celui-ci (les valeurs sont homogènes au sein du voisinage). La variabilité d'un instant d'une série S_l au sein de son voisinage est la valeur prise par le profil variance.

Ainsi, ${\sigma^M}_l[i]$ est la variabilité locale d'un instant. Une valeur ${\sigma^M}_l[i]$ forte correspond à un voisinage étalé autour de sa moyenne pour l'instant i, une valeur faible correspond à un voisinage compact.

Ainsi, nous distinguons la variabilité portée au sein du voisinage des instants, de celle portée entre les instants. Ceci nous incite à favoriser les instants présentant une faible variabilité. Nous allons à présent nous intéresser plus particulièrement à une structure de voisinage associée à un découpage des séries en une partition. Dans un premier temps, nous étudions les liens entre les instants au sein des séries d'une même classe.

1.2.b Compacité du voisinage des instants

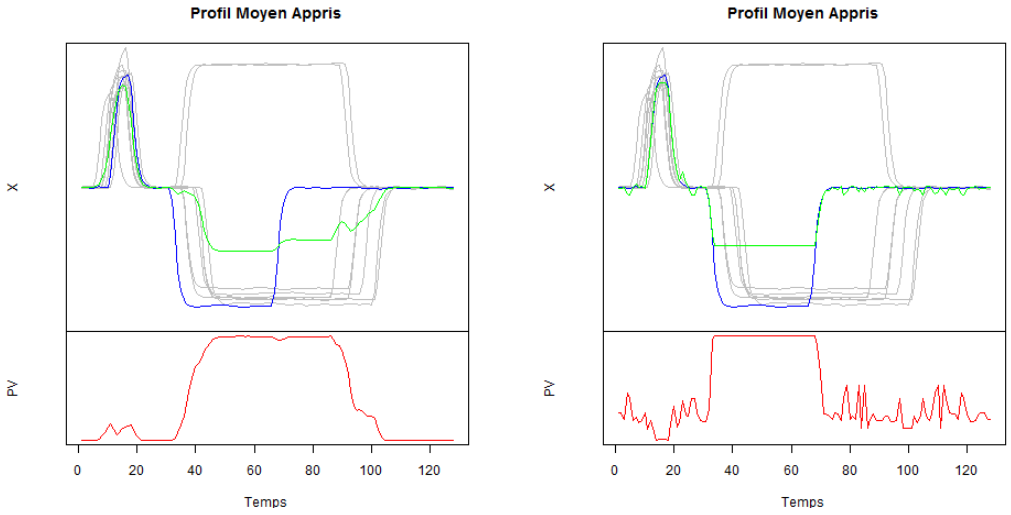
La variabilité d'une série S_l au sein de sa classe, qu'on note ${\sigma^+}_l$, se définit comme la variance du voisinage intra d'un instant :

$$ {\sigma^+}_l[i] = \frac{1}{n_k} \sum_{y_{l'}=y_l} \sum_{j=1}^{l} tMW_{ij}^{ll'} \left(S_j^{l'} - {\mathbb{S}^W}_i^{\,l} \right)^2 \tag{68} $$

où $\mathbb{S}^W$ est le profil moyen de S_l calculé sur la base du voisinage W.

$$\mathbb{S}^{W^l}[i] = \sum_{y_{l'}=y_l} \sum_{j=1}^{l} tW_{ij}^{ll'} S_j^{l'} \tag{69}$$

Les instants les plus caractéristiques sont donc ceux qui présentent une faible contribution à la variabilité intra-instants d'une série. En ce sens, la variance intra sera plus faible si les observations au sein d'un voisinage sont proches de la moyenne de voisinage.



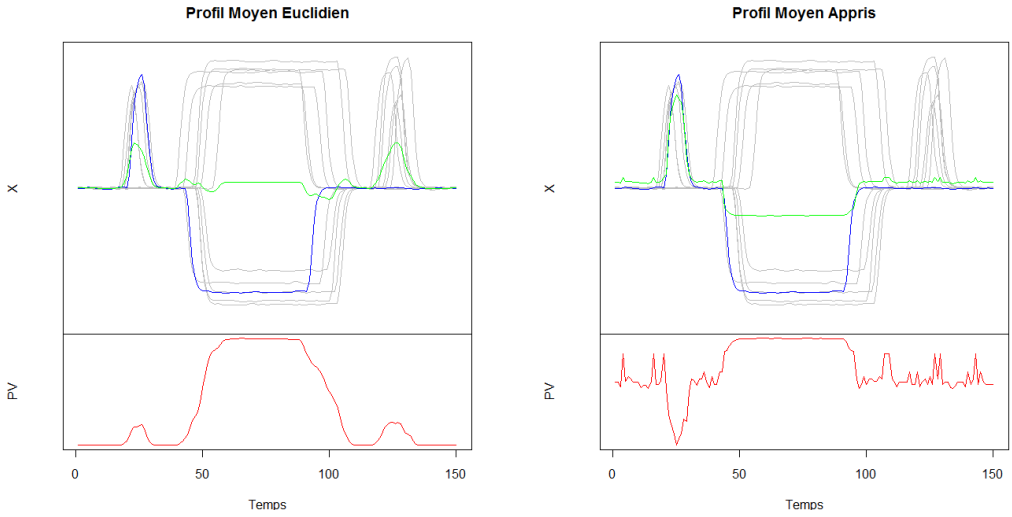
(a) Appariement euclidien (b) Appariement appris

FIGURE 36 – Profil variance associé à une série de la classe Begin (jeu BME)

La figure 36 donne en rouge le profil variance associé à un appariement euclidien (respectivement à l'appariement appris). Nous voyons sur la première figure que la variabilité est très faible sur les zones constantes situées en dehors des bosses, qu'elle est moyenne à hauteur de l'emplacement de la première bosse, et très forte autour du plateau central, du fait d'une forte variabilité entre les plateaux positifs et les plateaux négatifs. Dans le cadre des appariements appris (seconde figure), nous pouvons remarquer que la variabilité autour de la petite bosse est très faible. L'apprentissage a désigné la petite bosse comme l'élément caractéristique de la série.

La figure 37 concerne le jeu UMD. Nous voyons sur la première figure que la variabilité est très faible sur les zones constantes situées en dehors des bosses, qu'elle est moyenne à hauteur de l'emplacement de la première bosse, et très forte autour du plateau central, du fait d'une forte variabilité entre les plateaux positifs et les plateaux négatifs. Dans le cadre des appariements appris (seconde figure), nous pouvons remarquer que la variabilité autour de la petite bosse est très faible. L'apprentissage a désigné la petite bosse comme l'élément caractéristique de la série.

La variance des voisinages intra permet de mesurer le caractère caractéristique des instants. Cependant, certains instants caractéristiques sont communs à plusieurs classes. Nous allons donc chercher à évaluer le caractère différentiel de chaque instant.



(a) Appariement euclidien (b) Appariement appris

FIGURE 37 – Profil variance associé à une série de la classe Up (jeu UMD)

1.2.c Séparabilité des classes

Comme dans le cadre de la variabilité au sein des classes, la variabilité entre les classes est en général évaluée sur la base de l'ensemble des séries. La variabilité intra-instants entre les classes, qu'on note σ_l^-, se définit alors comme la variance du voisinage inter d'un instant.

$$\sigma^{-2^l}[i] = \frac{1}{\sum n_k} \sum_{C_k != y_l} \sum_{y_{l'}=C_k} \sum_{j=1}^{l} tMB_{ij}^{ll'} \left(S^{l'}[j] - \mathbb{S}^{B^l}[i] \right)^2 \tag{70}$$

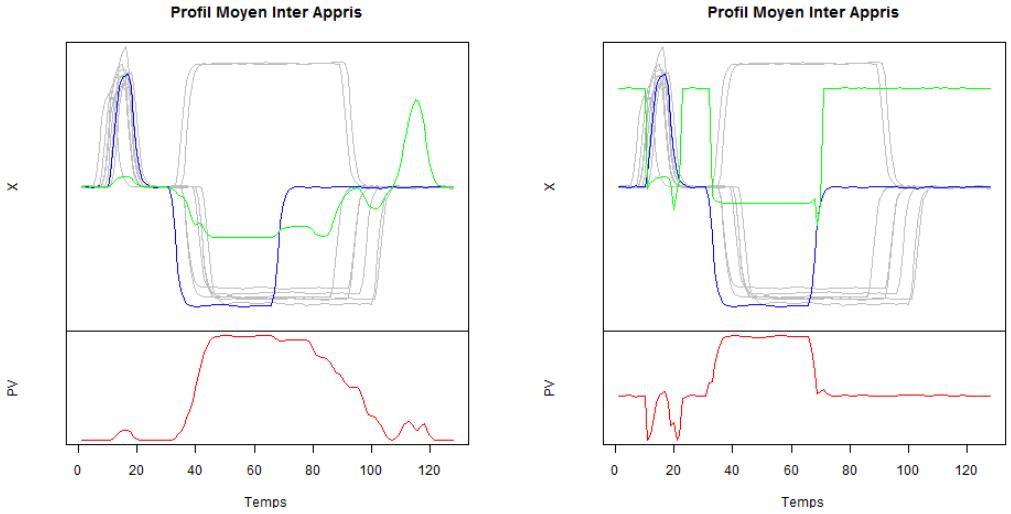
où $\mathbb{S}^{B^l}$ est le profil moyen inter de S^l.

A nouveau, une variabilité inter faible d'un instant signifie qu'il existe un ensemble homogène de valeurs au sein de toutes les séries de la classe, qui maximise l'écart entre une valeur observée et la moyenne sur ces valeurs, une valeur forte faisant état d'un ensemble très hétérogène. Les instants les plus différentiels sont donc ceux qui présentent à la fois une faible contribution à la variance inter-classes, et une faible contribution à la variabilité intra-instants.

La figure 38 donne en rouge le profil variance associé à un appariement euclidien entre les séries de classes différentes. Nous voyons sur cette figure que la variabilité se comporte de manière identique au cadre intra-classe : très faible sur les zones constantes situées en dehors des bosses, moyenne à hauteur des emplacements potentiels des deux petites bosses, et plus forte aux instants du plateau central, du fait d'une forte variabilité entre les plateaux positifs et les plateaux négatifs.

La définition de l'aspect discriminant des instants repose ainsi à la fois sur le profil moyen et sur le profil variance.

A partir de la notion de profils-moyen et de profils-variance, nous sommes à présent en mesure d'évaluer le caractère discriminant de chaque instant.



(a) Appariement euclidien (b) Appariement appris

FIGURE 38 – Profil variance associé la structure inter d'une série Begin

2 Instants discriminants associés à la variabilité au sein d'un voisinage

Nous définissons ainsi dans cette section des poids caractéristiques et différentiels associés à chaque instant, sur la base des profils-moyen et des profils-variance.

2.1 Instants caractéristiques

Nous considérons qu'un instant d'une série correspond à un événement caractéristique d'une structure de voisinage intra-classe, lorsqu'il trouve écho au sein des séries de sa classe. Nous définissons donc les instants caractéristiques comme étant les instants où existe une forte homogénéité au sein du voisinage et qui sont représentatifs de la série. En cet instant, le profil moyen est proche de la série initiale et le profil variance est faible. Dans le cadre d'un voisinage euclidien classique, il s'agit d'un événement qui se produit à un instant donné pour la majorité des séries. Pour la DTW, il s'agit d'un instant central que la recherche du chemin optimal tend à préserver lors de la déformation. Dans le cadre de la structure de voisinage apprise aux chapitres précédents, il s'agit d'un instant qui trouve des liens avec des instants de toutes les séries de la classe, où demeure une forte homogénéité. Nous proposons alors une pondération des instants caractéristiques sur la base de leur variabilité.

Aspect caractéristique d'un instant au sein de sa classe Nous considérons un instant d'une série caractéristique, si le profil moyen est peu déformé par rapport à la série initiale, et si la variance associée au voisinage de cet instant est faible.

Le poids associé à un événement caractéristique est défini de la manière suivante :

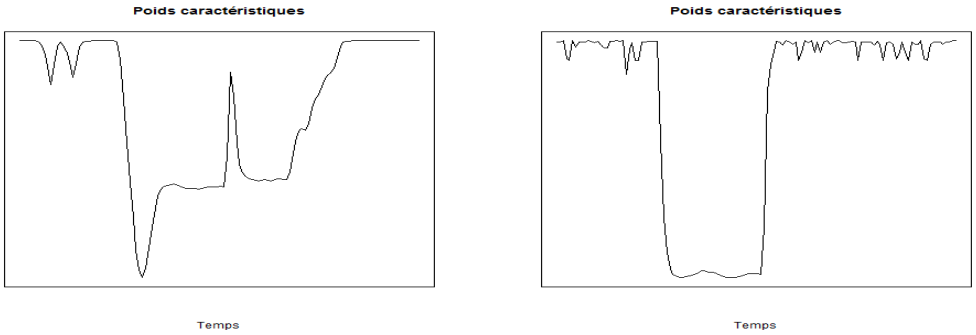
$$\alpha_W[i,l] = |S_l[i] - \mathbb{S}^{lW}[i]| \times \left(\sqrt{\sigma_l^+}\right) \tag{71}$$

Le terme α atteint son minimum lorsque le profil moyen coincïde avec la série initiale et que la variance est nulle.

Pour se ramener à un système de poids, on définit

$$P_W = \frac{e^{-\alpha_W[i,l]}}{sum(e^{-\alpha_W[i,l]})} \tag{72}$$

Le choix de passer à l'exponentielle pour exprimer les poids permet d'assurer à la fois la positivité et nous assure d'obtenir des poids bornés. Le poids est maximal pour les instants [i,l] dont le terme $\alpha_W[i,l]$ est minimal, c'est-à-dire ayant une variabilité faible.



(a) Appariement euclidien (b) Appariement appris

FIGURE 39 – Poids caractéristique associé à une série Begin

La figure 39 indique que la bosse centrale n'est caractéristique ni pour l'approche euclidienne, ni pour l'approche apprentissage. En effet, au sein des classes, la bosse centrale est tantôt orientée vers le haut, tantôt orientée vers le bas. Le reste de la série est plutôt caractéristique de la classe. Toutes les séries partagent la petite bosse au départ, ainsi que les zones constantes entre les bosses.

Nous obtenons ainsi un poids caractéristique de sa classe pour chaque instant. Nous définissons à présent de manière similaire des poids différentiels.

2.2 Instants différentiels

Nous considérons qu'un instant d'une série correspond à un événement différentiel d'une structure de voisinage inter-classes, lorsqu'il particularise la série vis-à-vis des séries des autres classes. Nous définissons les instants différentiels comme des instants où nous observons un écart fort entre la valeur observée et les valeurs prises au sein du voisinage. Pour qu'un instant soit différentiel et permette de distinguer nettement les séries, nous ne considérons comme différentiels que des instants pour lesquels le voisinage varie peu.

Dans le cadre d'un voisinage euclidien classique, il s'agit d'un instant pour lequel les séries n'appartenant pas à la même classe que la série étudiée prennent toutes une valeur semblable, et éloignée de celle de la série. Pour la DTW, il s'agit d'un instant ne s'alignant pas avec les séries des autres classes lors de la déformation. Dans le cadre de la structure de voisinage apprise aux chapitres précédents, il s'agit d'un instant qui se connecte avec certains instants d'autres classes desquels il s'éloigne fortement, mais où demeure une forte homogénéité. Nous proposons alors une pondération des instants différentiels sur la base de l'écart observé.

Caractère différentiel d'un instant avec les autres classes Nous considérons un instant d'une série comme étant distinctif, si, d'une part l'écart observé entre ce voisin et la

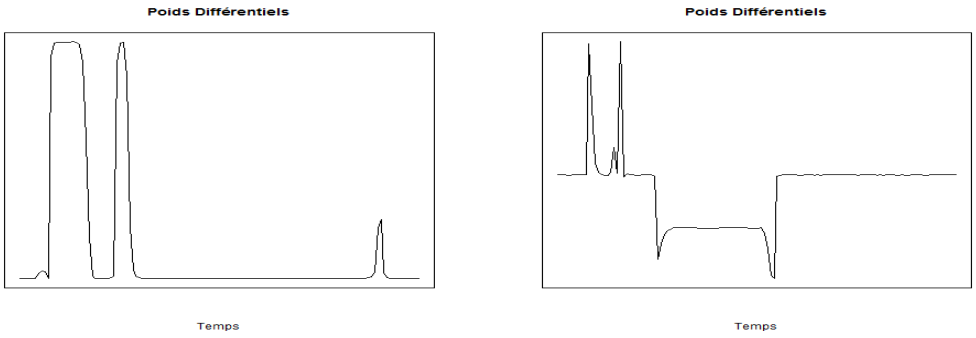
moyenne inter observée sur son voisinage est fort, et si, d'autre part, le voisinage est homogène autour du profil moyen.

Le poids associé à un événement différentiel est défini par :

$$\alpha_B[i,l] = \frac{\sigma^{-l}[i]}{abs\left(S_l[i] - \mathbb{S}^{lB}[i]\right)} \tag{73}$$

Pour se ramener à un système de poids, on définit

$$P_B = \frac{e^{-\alpha_B[i,l]}}{sum(e^{-\alpha_B[i,l]})} \tag{74}$$



(a) Appariement euclidien (b) Appariement appris

FIGURE 40 – Poids différentiel associé à une série Begin

L'approche euclidienne et l'approche apprentissage donnent un poids différentiel maximal à la première bosse, qui est en effet singulière de la classe.

Les instants les plus intéressants en vue de la discrimination d'un ensemble de séries temporelles sont les instants qui correspondent à des événements caractérisant les séries de la classe, mais également, qui permettent de distinguer les classes entre elles. Dans les jeux BME et UMD, les instants des plateaux sont caractéristiques, mais communs à chaque classe. Ils ne permettent pas de discriminer les classes. Les instants discriminants sont les instants à la fois caractéristiques et discriminants. Nous proposons dans la suite un système de poids discriminants associés aux instants des séries.

2.3 Instants discriminants

On appelle événement discriminant, pour un voisinage donné, un événement à la fois différentiel et caractéristique. Il correspond à un instant où les valeurs observées sont assez homogènes au sein de la classe et bien séparées de celles des autres classes. Nous proposons alors une pondération des instants discriminants.

Caractère discriminant d'un instant Nous considérons un instant d'une série étant discriminant, s'il présente un écart fort avec les valeurs moyennes des classes voisines, et si les valeurs moyennes sont représentatives, à travers une faible variabilité. Le poids associé à

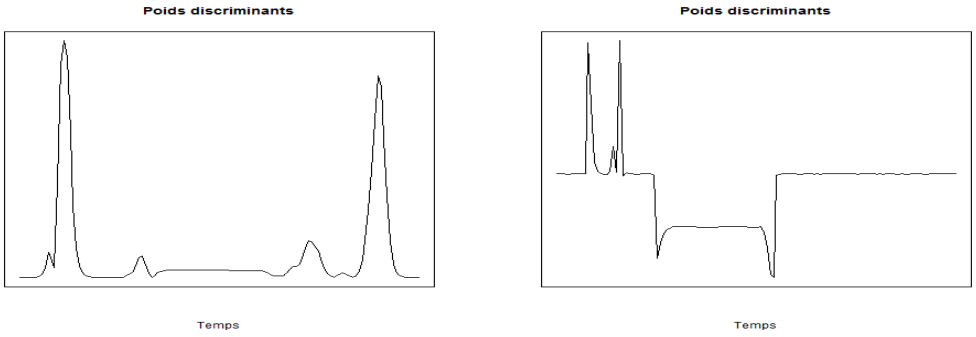
un événement discriminant est défini de la manière suivante :

$$\alpha_{\Delta} = \frac{abs\left(\bar{S}_{l_{MW|_{kppv}}}[i] - \bar{S}_{l_{MB|_{Imposteurs}}}[i]\right)}{\sigma^{+l}[i] \times \sigma^{-l}[i]} \tag{75}$$

Pour se ramener à un système de poids, on définit

$$P_{\Delta} = \frac{e^{-\alpha_{\Delta}[i,l]}}{sum(e^{-\alpha_{\Delta}[i,l]})} \tag{76}$$

Notons en particulier que si une série ne possède pas d'imposteurs, son poids discriminant est égal à son poids caractéristique. Le poids est maximal pour les instants [i,l] dont le terme $\alpha_{\Delta}[i,l]$ est minimal, c'est-à-dire ayant une variabilité faible, à la fois sur le profil moyen intra et sur le profil moyen inter, et dont les profils moyens s'écartent l'un de l'autre en cet instant.



(a) Appariement euclidien (b) Appariement appris

FIGURE 41 – Profil variance associé à une série

L'approche euclidienne a bien réussi à faire ressortir les instants discriminants, qui correspondent aux positions des deux plateaux. L'apprentissage se limite à reconnaître la première bosse, qui permet seule de discriminer les séries.

L'approche proposée ci-dessus repose sur la notion de profil moyen. La variabilité de la structure de voisinage retenue semble donner de bons résultats en matière de caractérisation des classes. Cependant, la notion de profil moyen différentiel est moins naturelle. De plus, pour les structures classiques, le fait d'avoir un nombre équilibré d'arêtes rend la variabilité comparable au sein des structures. En revanche, lorsque les structures de voisinage associées à chaque instant sont déséquilibrées, alors la variance peut être moins représentative. Nous proposons alors dans la suite une autre manière de définir des poids discriminants pour un ensemble de séries temporelles, fondée sur la notion d'entropie.

3 Définition de poids discriminants à partir de l'entropie d'une structure de voisinage

Présentation sommaire de l'entropie de Shannon L'entropie de Shannon est une grandeur utilisée dans de nombreux domaines allant de la théorie de l'information à la génétique. C'est une grandeur qui mesure l'écart d'une distribution à une distribution uniforme.

Définition 60 : (Entropie d'un systeme de poids)
L'entropie se définit comme une fonction $H : P = (p_1, ..., p_n) \mapsto -\sum_{i=1}^{n} p_i log(p_i)$.
L'entropie est une grandeur positive, vérifiant $0 \leq H(P) \leq log(n)$

Proposition 61 : (Quelques propriétés élémentaires de l'entropie de Shannon)

1. $\forall P = (p_1, \ldots, p_n), H(P) \geq 0$ *avec égalité si et seulement si* $\exists i_0 \backslash p_{i_0} = 1$
2. $\forall i$, $H(p_1, \ldots, p_n)$ *est continue en* p_i
3. $\forall \sigma \in S_n$ *(groupe des permutation),* $H(p_1, ..., p_n) = H(p_\sigma(1), ..., p_\sigma(n))$
4. $H(p_1, ..., p_n) \leq H(\frac{1}{n}, ..., \frac{1}{n}) = log(n)$. *La loi uniforme maximise l'entropie.*
5. $H(p_1, ..., p_n) = H(p_1 + p_2, p_3, ..., p_n) + (p_1 + p_2)(H(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}))$
6. $H(p_1, ..., p_n)$ *est une fonction concave :* $\forall \lambda \in [0, 1]$ $H(\lambda p_1 + (1-\lambda)q_1, ..., \lambda p_n + (1-\lambda)q_n) \leq \lambda H(p_1, ..., p_n) + (1-\lambda)H(q_1, ..., q_n)$

Une mesure de la singularité fondée sur l'entropie de Shannon Définissons la fonction $\mathbb{H}(P)$.

$$\mathbb{H}(P) = log(n+1) - H(P) \tag{77}$$

Cet indice est utilisé fréquemment pour mesurer la biodiversité. Des propriétés de l'entropie de Shannon, découlent les propriétés suivantes pour l'indice de diversité.

Proposition 62 :

1. *d'après 4* $\forall P = (p_1, \ldots, p_n), \mathbb{H}(P) > 0$
2. *d'après 2* $\forall i$, $\mathbb{H}(p_1, \ldots, p_n)$ *est continue en* p_i
3. *d'après 3* $\forall \sigma \in S_n$ *(groupe des permutation),* $\mathbb{H}(p_1, ..., p_n) = \mathbb{H}(p_\sigma(1), ..., p_\sigma(n))$
4. *d'après 4 et 1* $\mathbb{H}(p_1, ..., p_n) \geq \mathbb{H}(\frac{1}{n}, ..., \frac{1}{n}) = log(1 + \frac{1}{n})$. *La loi uniforme minimise la fonction* $\mathbb{H}$.

L'entropie est une manière d'étudier l'écart d'une distribution de probabilités à l'uniforme. L'entropie est une fonction qui atteint son minimum pour une distribution uniforme, et son maximum pour une distribution unimodale. Nous allons utiliser l'entropie pour la définition de poids caractéristiques. Nous considérons en effet qu'un instant saillant trouvant écho parmi les instants des séries de sa classe, correspond à un événement plus caractéristique de la classe.

3.1 Entropie d'un instant au sein d'un couple de séries liées par une structure de voisinage

Dans les approches que nous avons adoptées, la normalisation des matrices fluctue. Si nous renormalisons dans le cadre booléen les blocs matriciels, alors, nous obtenons pour chaque

instant une distribution de probabilité des poids $MW_{ii'}^{ll'}$ (respectivement $MB_{ii'}^{ll'}$). Notons $\mathbb{P}$ cette probabilité.

– Dans le cas booléen, tous les termes non nuls d'une ligne de la matrice définissant les poids des arêtes entre deux séries sont égaux. Ainsi, $\mathbb{P}_{i'} \in \{0, \frac{1}{\#(MW_{ii'}^{ll'} \neq 0)}\}$

$$\mathbb{H}(\mathbb{P}) = log(n+1) + \frac{\#(MW_{ii'}^{ll'} \neq 0)}{\#(MW_{ii'}^{ll'} \neq 0)} log(\frac{1}{\#(MW_{ii'}^{ll'} \neq 0)}) \quad (78)$$

$$= log(n+1) - log(\frac{1}{\#(MW_{ii'}^{ll'} \neq 0)}) \quad (79)$$

$$= log(\frac{n+1}{\#(MW_{ii'}^{ll'} \neq 0)}) \quad (80)$$

$$(81)$$

S'il y a un unique lien entre l'instant i de S^l et les instants de $S^{l'}$, alors $\mathbb{H}(\mathbb{P}) = log(n+1)$; si tous les liens sont présents, $\mathbb{H}(\mathbb{P}) = log(1 + \frac{1}{n})$.

– Dans le cas progressif, les poids des arêtes sont quelconques. Par continuité et concavité de la fonction $\mathbb{H}$, plus le voisinage est dense, plus la fonction sera faible.

$$(82)$$

La matrice des instants indique les liens possibles entre les instants de la série S^l et ceux de la série $S^{l'}$. Toutes ces arêtes sont celles qui minimisent la variance intra ou maximisent la variance inter. Cependant, un événement de S^l peut se lier à plusieurs instants de la série $S^{l'}$. Ainsi, plus la structure de voisinage associée à un instant est compacte, moins il y a d'ambiguïté sur le choix des instants. Les instants les plus discriminants sont ceux pour lesquels la structure de voisinage s'éloigne le plus de la structure uniforme. Un instant est discriminant s'il se connecte à peu d'arêtes au sein du voisinage. A partir de la notion d'entropie appliquée dans le cadre d'un ensemble de séries, nous définissons pour chaque série un profil entropique associé à chaque série.

3.2 Profil entropique d'une paire de séries liées par une relation de contiguïté

Soit S_l et $S_{l'}$ deux séries temporelles, et $M^{ll'}$ la structure de contiguïté qui lie les instants de S_l avec ceux de $S_{l'}$, i.e., $M_{ii'}^{ll'} = 1$ si et seulement si les instants i et i' sont liés. La matrice $M^{ll'}$ est symétrisée en une matrice sM avec $sM_{ii'}^{ll'} = \frac{M_{ii'}}{\sum_{k=1}^{T} M_{ik}^{ll'}} + \frac{M_{i'i}^{ll'}}{\sum_{k=1}^{T} M_{ki'}^{ll'}}$. En outre, sM vérifie la relation suivante

$$\forall i \in \{1, \ldots, T\}, \ \sum_{k=1}^{T} sM_{ik}^{ll'} = 1$$

Nous pouvons donc, pour chaque instant i, calculer $\mathbb{H}^{ll'}(i) = log(n+1) + \sum_{k=1}^{T} sM_{ik}^{ll'} log(sM_{ik}^{ll'})$. Nous obtenons un profil entropique pour le couple de séries $S_l, S_{l'}$ On définit alors le poids entropique caractéristique d'une série S_l, soit ${P_W}^l$.

$$\beta_W{}^l[i] = \prod_{S_{l'}|y_{l'}=y_l} \mathbb{H}^{ll'}[i]^{1/n_{y_l}} \tag{83}$$

$$P_W{}^l[i] = \beta^l_W[i]/sum_j(\beta^l_W[j]) \tag{84}$$

Pour tout i, ${\beta_W}^l_i$ est égal à la moyenne géométrique des profils entropiques associés à la série S_l pour toutes les séries S'_l dans la classe de S_l. La moyenne géométrique est moins sensible que la moyenne arithmétique aux valeurs les plus élevées d'une série de données. De plus, elle a tendance à pénaliser plus fortement les valeurs faibles. Si à un instant est associée une distribution proche de l'uniforme pour certaines séries, alors cet instant doit avoir globalement un poids faible. Le vecteur β_W est alors normalisé pour se ramener à une distribution de probabilités. De même, on définit le poids entropique différentiel de la série S_l ${p_B}^l$ comme une moyenne géométrique normalisée des profils entropiques associés à la série S_l pour toutes les séries S'_l qui ne sont pas dans la classe de S_l.

Définition 63 : (Poids entropique d'une série)

$$\forall i \in \{1,\ldots,T\},\ {p_W}^l_i = \frac{\prod_{l'\setminus y_l=y_{l'}} \mathbb{H}^{ll'}(i)}{\sum_t(\prod \mathbb{H}^{ll'}(t))} \tag{85}$$

$$\forall i \in \{1,\ldots,T\},\ {p_B}^l_i = \frac{\prod_{l'\setminus y_l\neq y_{l'}} \mathbb{H}^{ll'}(i)}{\sum_t(\prod \mathbb{H}^{ll'}(t))} \tag{86}$$

La notion de profil entropique est liée à l'existence de structures de voisinage non uniformes en ce qui concerne le nombre d'instants, entre les différentes séries. Comme les poids entropiques associés à un appariement euclidien sont uniformes, dans cette partie, nous ne considérerons que l'appariement appris dans le cadre de l'apprentissage booléen.

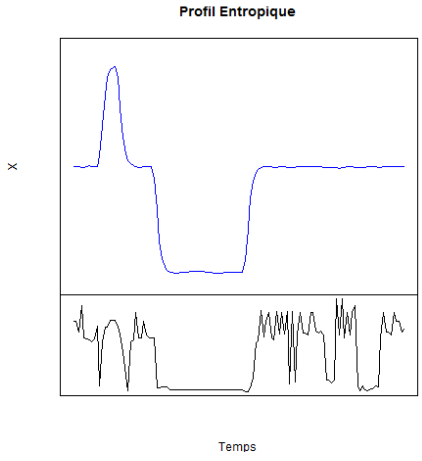


FIGURE 42 – Poids entropiques associés à une série Begin

Les profils entropiques associés à une structure de contiguïté permettent de mesurer le degré de fiabilité de l'appariement appris pour faire se correspondre des événements semblables. Dans le cas où la structure de contiguïté est associée à une structure de partition,

nous souhaitons définir des profils entropiques qui soient discriminants. Nous allons donc, à partir de la définition associée à une structure de contiguïté quelconque définie ci-dessus, introduire des profils caractéristiques et différentiels, et les rassembler pour obtenir des profils discriminants.

3.3 Silhouette entropique discriminante d'une classe de série

Nous cherchons à définir pour chaque classe une silhouette entropique qui soit caractéristique pour les séries qui composent la classe et différentielle vis-à-vis des séries des autres classes. Nous définissons donc dans un premier temps la silhouette entropique caractéristique (respectivement différentielle) d'une classe, comme moyenne arithmétique des profils caractéristiques (respectivement différentiels) pour toutes les séries de la classe. Un instant d'une classe sera discriminant s'il se présente à la fois comme caractéristique des séries de sa classe, et comme différentiel au regard des séries des autres classes.

Les structures de voisinage apprises étant discriminantes, les profils entropiques appris sont discriminants. On définit ainsi la silhouette entropique discriminante d'une classe, comme un système de poids entropiques discriminants PED dont le poids de l'instant i est le suivant

$$PED_i = \left(\frac{1}{n_k} \sum_{l=1}^{n_k} (p^l_{W\,i} + p^l_{B\,i}) \right) \tag{87}$$

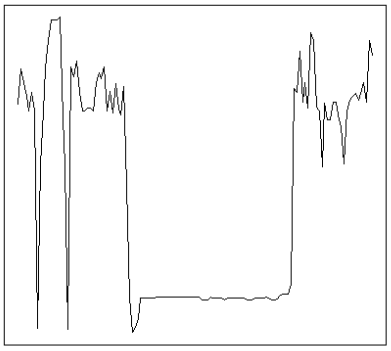


FIGURE 43 – Poids Entropique Discriminant PED pour une série de la classe Begin

Nous voyons que les profils discriminants obtenus mettent en lumière les événements les plus discriminants des séries pour les jeux simulés UMD et BME. Cependant, pour certains jeux de données, les séries d'une classe peuvent être définies par plusieurs profils. Par exemple, dans la classe Up, les petites bosses sont tantôt orientées vers le bas, tantôt orientées vers le haut. Les profils discriminants peuvent ainsi être plus fidèles s'ils ne sont pas construits à partir de l'ensemble des séries, mais seulement à partir des séries qui leur sont les plus proches.

4 Choix d'un sous-ensemble de séries en vue de la définition d'instants discriminants

Nous proposons dans cette section, d'affiner les profils discriminants, en limitant leur calcul à un sous-ensemble de séries proches de la série considérée.

4.1 Objectif

En fonction des objectifs visés, il est parfois important de limiter l'étude à certains couples de série. Lors d'une classification de type "plus proches voisins", une série test est affectée à la classe de la série qui lui est la plus proche. Ainsi, pour limiter les erreurs, l'approche classique consiste à séparer les classes au maximum. Pour ce faire, il s'agit d'une part de minimiser la compacité des classes, c'est-à-dire diminuer la variabilité des séries au sein de la classe et d'autre part, de maximiser la séparabilité des classes, i.e., augmenter l'écart entre séries de classes différentes c'est-à-dire augmenter la variabilité entre les classes de séries.

Supposons connue une mesure de proximité entre séries temporelles, notée D. D peut être par exemple une distance usuelle à l'instar de la Distance euclidienne ou de la DTW. Nous allons voir comment le choix d'un sous-ensemble de séries permet d'affiner les critères de compacité et de séparabilité des classes. L'apprentissage des blocs inter se faisant de manière séparée, sur la base d'une métrique prédéfinie, nous pouvons limiter notre étude à la classe la plus proche. Toutefois, ceci peut se faire au détriment d'autres classes proches elles aussi. Un instant peut être discriminant pour un couple de classes, mais pas discriminant globalement. Ce qui importe véritablement pour classer correctement une série est de la rapprocher au sein de la classe qui lui ressemble et d'éloigner d'elle toute série qui potentiellement, pourrait entraîner une confusion avec elle.

Le fait de calculer le profil moyen sur la base de toutes les séries, toutes classes confondues, donne autant de poids, dans le calcul, aux séries des classes ayant un profil proche de la classe de S_l, qu'à celles qui sont lointaines. L'objectif est de tirer un poids différentiel de l'apprentissage des poids de voisinage, en vue de classifier au mieux les séries. De fait, il est plus intéressant de chercher à différencier les séries appartenant aux classes qui se ressemblent. Sur la base de la distance D définie précédemment, nous cherchons la classe la plus proche :

$$ppcl[S_l] = argmin_{y_k} \sum_{y_{l'}=C_k} D(S_l, S_{l'}) \tag{88}$$

La variance inter peut alors se calculer sur la base des séries de cette classe.

$$\sigma_l^-[i] = \frac{1}{\sum n_k} \sum_{y_{l'}=ppcl[S_l]} \sum_{j=1}^{l} tMB_{ij}^{ll'} \left(S_{l'}[j] - \bar{S}^B[i]\right)^2 \tag{89}$$

où $\bar{S}^B$ est le profil moyen de S.

Des profils différents peuvent apparaître au sein d'une même classe. Dans le cadre d'une approche 1NN, il peut s'avérer important de se limiter à un ensemble réduit de séries. De la même façon, les instants distinctifs n'ont d'intérêt que pour séparer des séries proches. Nous apprenons donc pour une série les instants caractéristiques sur la base des séries qui lui sont les plus proches au sein de sa classe, et les instants distinctifs sur la base des séries les plus proches dans les autres classes. Nous définirons en particulier la notion d'imposteurs.

Compacité de la classe Le choix de considérer toutes les séries peut poser problème, notamment quand il y a plusieurs profils dans la classe. L'objectif de la classification, du fait du choix de l'approche "plus proches voisins", consiste à rapprocher une série de ses voisines les plus proches, et il n'est pas nécessaire que les classes soient les plus compactes possible, au sens général du terme. En effet, il importe seulement de rapprocher chaque série de ses voisines les plus proches au sens de la mesure de proximité D. Nous pouvons alors limiter le calcul de la variabilité d'une série temporelle aux séries qui lui sont les plus proches. Nous pouvons, sur la base de la distance D, définir pour chaque série un voisinage constitué des k séries dans la classe des plus proches au sens de la distance D. On définit la matrice de voisinage $MW|_{kppv}$ de la manière suivante :

$$(S)\begin{cases} \text{si } S'_l \text{ est dans les séries voisines de } S_l MW|^{ll'}_{kppv} & = MW^{ll'} \\ \text{sinon} MW|_{kppv} & = 0 \end{cases}$$

$$\sigma_l^+[i] = \frac{1}{n_k} \sum_{\substack{y_{l'}=y_l \\ S_{l'} \text{voisine de } S}} \sum_{j=1}^{l} tMW_{ij}^{ll'} \left(S_{l'}[j] - \bar{S}_{lMW|_{kppv}}[i] \right)^2 \tag{90}$$

où $\bar{S}^W_{lMW|_{kppv}}$ est le profil moyen de S_l calculé sur la base des k plus proches voisines de S_l.

Pour de nombreux problèmes de discrimination, en particulier la classification d'une série au sein d'une partition de séries temporelles, il est plus utile de rechercher un critère de différenciation permettant de distinguer, pour une série donnée, les séries qui lui sont voisines et ne sont pas de sa classe, que de différencier toutes les séries entre elles.

4.2 Séparabilité des classes : Notion d'imposteur

Reprenons la notion d'imposteur introduite par Weinberger *et al.* (2006). On appelle imposteur toute série qui se situe dans la sphère des plus proches voisines de S_l. On définit tout d'abord le nombre de voisine k au sein de la classe qui nous intéresse.

Définition 64 : (Imposteur)
Une série Simp est un imposteur de S si $d(S, Simp) < max(d(S, S_i)|i \in k - ppvdeS)$

Un imposteur d'une série S se définit alors comme un individu dont la distance à la série S est plus faible que certains de ses plus proches voisins. Cette série présente un risque plus important d'induire une erreur de classification. L'objectif est donc de chercher une métrique qui éloigne les imposteurs de la série de départ. Ceci peut faire l'objet d'un apprentissage. Nous recherchons une métrique qui sépare les imposteurs. A partir de cette métrique, se redéfinit une matrice de dissimilarité, de laquelle nous pouvons extraire une nouvelle structure d'imposteurs.

Ainsi, la définition des profils discriminants proposés se généralise à un sous-ensemble de séries, permettant, en fonction des besoins, d'affiner les profils. Nous allons à présent montrer comment nous pouvons appliquer les appariements appris, ainsi que les profils discriminants qui en découlent.

5 Quelques applications des appariements appris

Les sorties du processus d'apprentissage intra et du processus d'apprentissage inter sont un couple de matrices d'appariements, où chaque bloc de l'appariement intra (respectivement inter) décrit l'appariement entre deux séries de la même classe (respectivement de deux classes différentes). Les objectifs de l'apprentissage des blocs discriminants, peuvent être de diverses natures (classification, définition d'un masque discriminant, lissage prenant en compte les éléments discriminants des classes...).
Les premières parties de ce chapitre ont permis d'extraire certaines applications, pour la définition d'un profil moyen ou de poids discriminants. L'entropie, présentée dans la partie précédente, vise à définir un masque décrivant toute la classe. Elle est fondée sur l'information contenue dans les blocs (au sens de Shannon). Dans certains cas, les appariements sont utilisés différemment pour définir une silhouette moyenne tenant compte des spécificités de chaque série liée à une série S_l. Selon les cas, les blocs appris peuvent aussi nécessiter certaines transformations préalables permettant d'extraire des informations plus spécifiques. Afin d'extraire une structure globale à toute la classe, il est judicieux d'effectuer une moyenne des appariements appris entre la série S_l et toutes les séries qui lui sont liées. Enfin, dans d'autres types d'application, nous cherchons à définir un alignement discriminant.

5.1 Utilisation des structures de voisinages apprises en vue de la définition d'un masque pour les séries

Nous avons obtenu pour chaque série un profil discriminant et une matrice d'appariement pour lier cette série aux autres. Ces informations permettent de mettre en lumière les événements les plus importants pour la discrimination des séries. Nous proposons plusieurs façons de construire de tels masques, soit directement à partir des matrices d'appariement apprises, soit à partir des profils discriminants associés.

5.1.a A partir des poids de voisinage appris

Les structures de voisinages apprises se définissent sous la forme de blocs matriciels. A chaque arête de couple $((i,i'),(l,l'))$ est affectée un poids. Nous pouvons considérer pour chaque série, la moyenne de ces blocs notée $Mmoy$. Cette moyenne peut être constituée des blocs discriminants intra, inter, ou de la totalité des blocs.

$$Mmoy^{l}_{ii'} = \sum_{l=1}^{n} M^{ll'}_{ii'} \tag{91}$$

- Dans le cas booléen, une arête est activée ou non. Ainsi, au sein d'un bloc $M^{ll'}$, le poids $M^{ll'}_{ii'}$ sera fort si l'instant i de la série S^l est connecté à l'instant i' de $S^{l'}$ et qu'il est connecté avec peu d'autres instants. Lorsque nous calculons la moyenne, le poids est lié au nombre de fois où les arêtes $\widehat{(i,i')}$ liées à la série S^l sont activées. Le bloc $Mmoy^l$ est normalisé en ligne.
- Dans le cas progressif, la ligne i de la matrice $M^{ll'}$ décrit une distribution de poids au sein des instants j de la série $S^{l'}$. La moyenne des blocs matriciels donne un bloc matriciel normalisé en ligne.

Ces masques sont intrinsèques à chaque série. Le terme général de la matrice correspond à la connectivité de l'instant i de la série S_l avec toutes les autres séries considérées. Ils renseignent sur l'importance des liens, et des régions au sein de la série qui sont connectées. Toutefois, lors du calcul de la moyenne, sont prises en considération des arêtes pouvant n'apparaître que dans peu de séries, mais avec un poids fort. Lors de la moyenne, elle peuvent prendre la même importance que des arêtes apparaissant avec un poids moyen dans l'ensemble des séries. La normalisation en ligne est préservée lors du calcul de la moyenne. De ce fait, peuvent être distinguées deux types d'arêtes, les arêtes ayant un poids inférieur au poids de la distribution uniforme (hypothèse agnostique de voisinage), et celles ayant un poids uniforme. A l'instar du seuillage apparaissant dans l'algorithme, nous pouvons "seuiller" les voisinages obtenus en ne conservant que les arêtes dont le poids est supérieur ou égal à l'uniforme, l'uniforme étant la limite nous assurant l'existence de tels poids. Les masques obtenus sont alors renormalisés pour se ramener à nouveau à un système de poids, soit par une transformation linéaire ramenant le minimum à 0 et le maximum à 1, puis une renormalisation, soit par renormalisation directement.

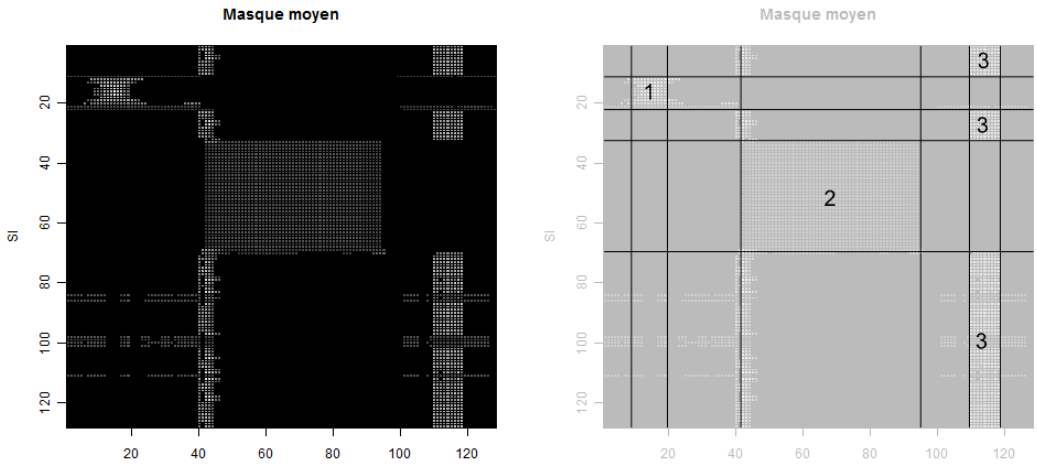


FIGURE 44 – Bloc moyen associé à une série Begin du jeu BME

La figure 44 décrit l'appariement moyen à une série particulière de la classe Begin. Nous remarquons une première zone claire (zone 1 sur le schéma de droite) correspondant aux couplages des petites bosses et une zone plus foncée entre les deux grandes bosses (zone 2). En effet, pour certaines séries, la bosse centrale apparaît. Elle conserve un poids moindre lors de l'appariement moyen. C'est dû au fait que la bosse centrale pour la série étudiée est orientée vers le bas, ce qui la distingue de la classe middle. La zone 3 est une zone discriminante, qui s'est dégagée lors du processus différentiel pour distinguer la classe Begin et la classe End.

La figure 45 décrit l'appariement moyen à une série particulière de la classe Up. Nous remarquons deux zones claires (zone 1 sur le schéma de droite) correspondant aux couplages de la petite bosse située en fin de série, couplées avec les petites bosses qui peuvent arriver au début et à la fin des séries de la classe. Une zone plus foncée lie les deux grandes bosses (zone 2), de la même façon que pour le jeu précédent. Elle conserve également un poids moindre lors de l'appariement moyen. La zone 3 correspond au couplage des grandes bosses avec les

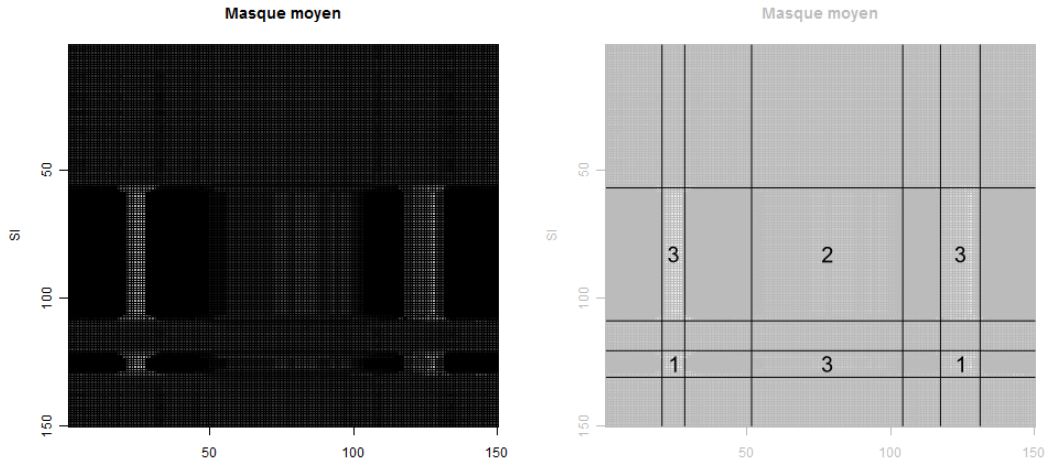


FIGURE 45 – Bloc moyen associé à une série Up du jeu UMD

petites, qui se couplent lorsque la grande bosse centrale est de même signe que la petite.

5.1.b A partir de l'entropie des instants

L'idée des masques entropiques est de calculer des masques discriminants en fonction de l'entropie des instants.

Nous lisons au sein des blocs deux types d'information, d'une part, sur les lignes, nous lisons la distribution des poids des arêtes conditionnellement à chaque instant, et en colonne, nous lisons l'attractivité des instants. Le poids de chaque ligne est égal. En revanche, le poids de chaque colonne varie au sein des séries. Plus un instant est lié, plus les colonnes ont un poids élevé.

Entropie des lignes L'entropie nous donne la précision du voisinage. Nous considérons que plus un instant s'éloigne du couplage complet, plus il est discriminant. En effet, si un instant est connecté à peu d'arêtes, il y a peu d'ambiguïté sur les couplages associant cet instant. Au contraire, si le couplage est proche de l'uniforme, il n'y a pas de préférences entre les arêtes.

Poids marginal des colonnes Au contraire des lignes, les poids des colonnes ne sont pas normalisées. Une ligne est discriminante si le nombre d'arêtes fortement connectées est réduit. Cela signifie que le choix des instants est net. En revanche, les colonnes qui nous intéressent sont celles qui ont le poids maximal. On définit alors un poids en colonne qui est égal à la somme des poids de chaque colonne. Une série aura un poids en colonne fort si des instants de S^l se connectent fortement avec lui.

Masques entropiques Nous définissons le masque entropique du couple de série $(S^l, S^{l'})$ comme la matrice $M_{Hii'} = PL_i * PC_j$

Nous observons sur la figure 48 les masques entropiques des séries précédentes. Les régions correspondant aux bosses ont un poids plus fort. Les masques entropiques ont permis de

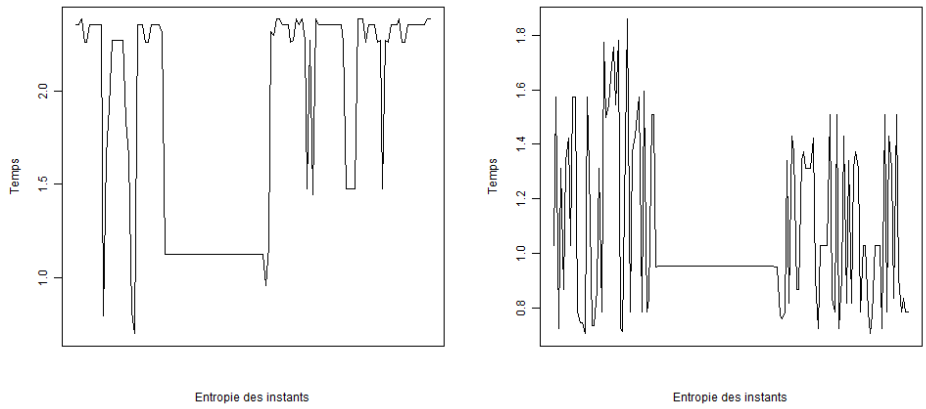
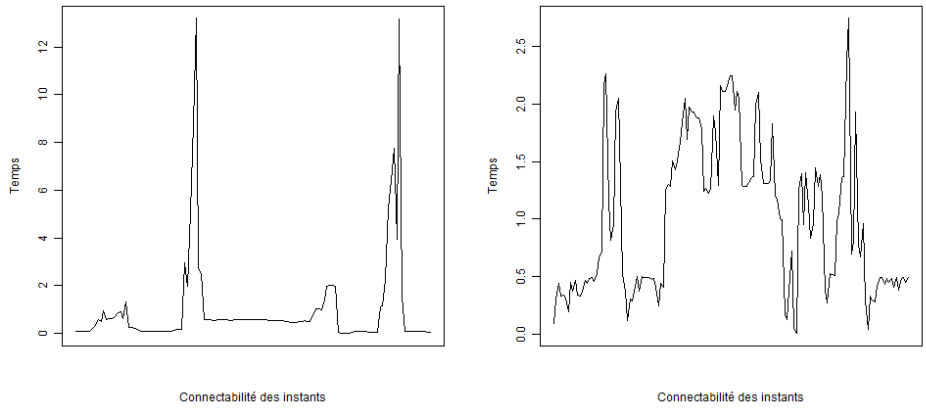
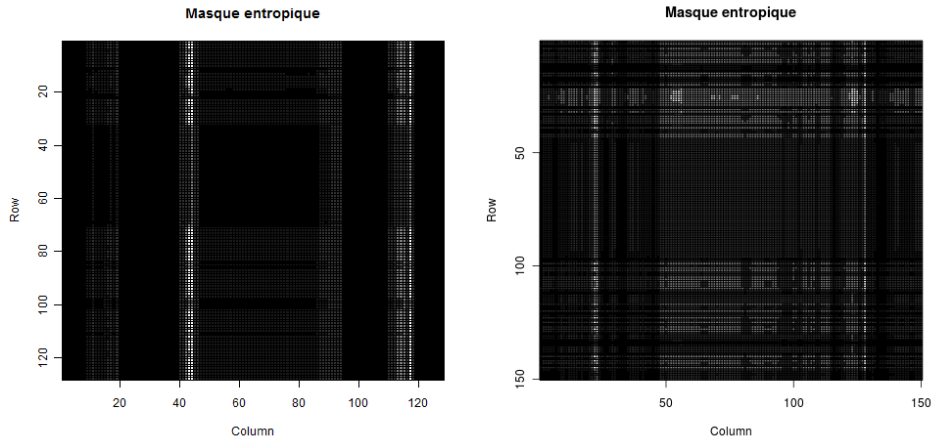


FIGURE 46 – Poids entropique des instants



(a) Begin (b) Up

FIGURE 47 – Poids marginal des instants



(a) Begin (b) Up

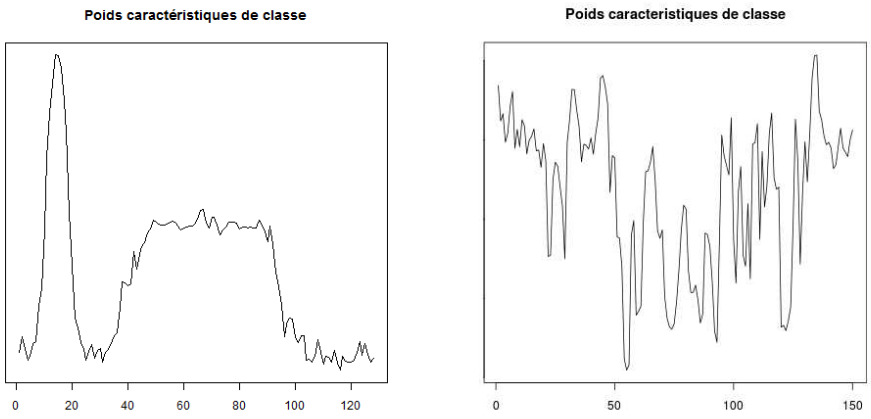
FIGURE 48 – Masques entropiques des séries

retrouver les événements discriminants du jeu BME. Pour en déduire une information plus générale à chaque classe de séries, nous souhaitons fusionner les masques de chaque série pour obtenir un profil commun à la classe.

5.2 Profil de classes

Les poids entropiques appris pour chaque série peuvent permettre de définir un profil de classe en rassemblant ces poids. Nous présentons ici la manière de définir un profil combinant les éléments caractéristiques d'une part, et les éléments discriminants.

Profil de classe caractéristique Le profil de classe caractéristique est la moyenne arithmétique des profils entropiques des séries. Un instant dont le poids est fort au sein du profil de classe est un instant apparaissant dans certaines séries. L'intérêt des profils de classe est de déterminer une région où peuvent apparaître potentiellement des événements importants. Nous déterminons ainsi les zones de fort potentiel ; ces zones doivent être considérées plus particulièrement lors de la comparaison de séries.



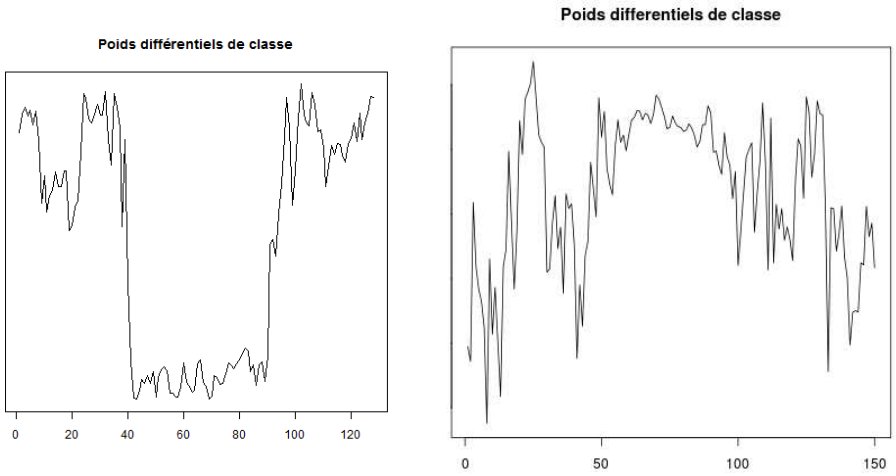
(a) Begin (b) Up

FIGURE 49 – Profil caractéristique des classes Begin et Up

Profil de classe discriminant On définit les profils différentiels entre paires de classes comme la moyenne des profils. Le profil de classe discriminant est une moyenne géométrique des profils différentiels entre paires de classes.

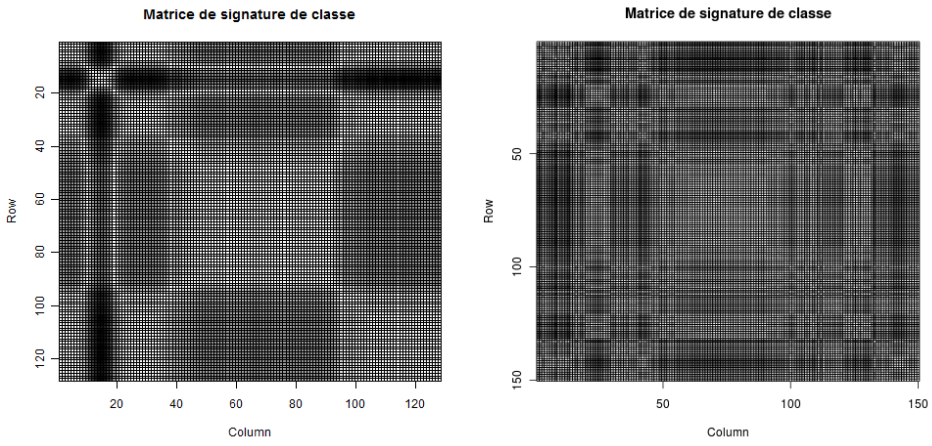
Signature de classe La signature de classe consiste à proposer un masque qui lie les instants dont les poids sont proches au sein de séries. Si des événements apparaissent dans les séries avec un poids fort, ils peuvent être liés. Le principe des poids de classe est d'offrir un moule dans lequel les instants qui sont liés d'un poids faible ont peu de chances de se lier. Cette signature de classe peut en particulier servir d'initialisation au processus d'apprentissage.

Nous avons extrait des matrices d'appariement une information discriminante pour chaque classe de séries. Nous allons à partir des appariements appris, des profils discriminants et des masques définis ci-dessus, définir plusieurs mesures de proximité qui soient discriminantes.



(a) Begin (b) Up

FIGURE 50 – Profil discriminant des classes Begin et Up



(a) Begin (b) Up

FIGURE 51 – Signature des classes Begin et Up

Nous évaluerons ensuite la performance de ces mesures de proximité pour la classification k-NN de séries issues des jeux simulés UMD et BME.

5.3 Définition de distances entre séries temporelles fondées sur les appariements appris

Nous voulons, à partir des appariements appris, définir des mesures de dissimilarité, en vue d'une classification supervisée. Nous avons vu deux types d'application des blocs appris, d'une part, l'utilisation des blocs pour définir une moyenne pondérée des arêtes, afin d'obtenir un profil moyen caractéristique et un profil moyen différentiel de toutes les séries temporelles. Nous avons également utilisé les appariements appris pour la définition de poids et de masques discriminants. Nous dégageons alors deux types d'approches pour la définition de tels indices de dissimilarité.

L'objectif, au sein de cette section, est de généraliser les métriques classiques à partir des appariements appris.

5.3.a Dissimilarités fondées sur le profil moyen

L'idée de ce type d'approche est de substituer à chaque série son profil moyen. En effet, le profil moyen est une reproduction de la série prenant en considération les éléments discriminants de la série au sein de sa classe. Un événement d'une série qui serait singulier au sein d'une classe sera lissé au cœur du profil moyen. Nous pouvons être également conduits à diminuer l'impact des instants pour lesquels le profil moyen s'éloigne de la valeur de la série. Nous allons alors, à partir du profil moyen, étendre les distances usuelles. Notons que les extensions que nous proposons aux distances classiques sont intrinsèques à chaque série.

Extension de la distance euclidienne La distance euclidienne est une mesure de dissimilarité fondée sur les écarts en valeurs observés entre paires d'instants. La distance euclidienne compare les séries instant par instant. Nous définissons alors une extension de la distance euclidienne en substituant dans la définition les valeurs prises par le profil moyen aux valeurs prises par la série.

$$S_1 = (x_1, x_2 ... x_n)$$
$$S_2 = (y_1, y_1 ... y_n)$$
Distance Euclidienne :

$$d_E^{PM[S_1]}(S_1, S_2) = \sqrt{\sum_{i=1}^{n} p_i^{S^1} \frac{\alpha^{S^1}}{(MW^{S^1} * X_i - S_{1i})} (MW^{S^1} * X_i - S_{2i})^2}$$
$$\text{avec } \alpha^{S^1} = 1/sum_i(\frac{1}{(MW^{S^1} * X_i - S_{1i})})$$

où MW^{S^1} correspond aux lignes représentant les instants de la série S^1, et $p_i^{S^1}$ est un poids discriminant. Le terme $\frac{1}{(MW^{S^1} * X_i - S_{1i})}$ correspond à l'écart entre la série et son profil moyen, pour favoriser les instants où le profil moyen. α^{S^1} est un terme de normalisation qui transforme ces écarts en un système de poids, assurant l'équité des différentes séries.

En particulier, cette distance assure, sur la base d'un alignement euclidien, plus d'importance aux instants discriminants.

Extension de la DTW à partir de masques L'algorithme de déformation temporelle dynamique (DTW) est une méthode qui recherche un appariement optimal entre deux séries temporelles, sous certaines restrictions. Les séries temporelles sont déformées pour coller au mieux l'une à l'autre. On appelle alignement un chemin au sein des instants vérifiant les propriétés suivantes.

Un alignement se caractérise par la donnée de deux vecteurs $u = (u_1, u_2, ...u_r)$ et $v = (v_1, v_2, ...v_r)$ tels que

- $u_1 = x_1, v_1 = y_1, u_r = x_n, v_r = y_n$
- si $(u_i, v_i) = (x_{k_1}, y_{k_2})$, alors

$$(u_{i+1}, v_{i+1}) \in \{(x_{k_1}, y_{k_2+1}), (x_{k_1+1}, y_{k_2}), (x_{k_1+1}, y_{k_2+1})\}$$

L'utilisation des masques dans la DTW permet de contraindre l'alignement à exister dans un certain voisinage temporel. Les masques classiques sont les bandes (par exemple bandes de Sakoe–Chiba, parallélogramme d'Itakura,...) qui limitent les alignements possibles à des régions situées autour de la diagonale, i.e., de l'alignement euclidien. Nous voulons définir des masques qui privilégient des connections entre des instants discriminants. Un instant discriminant doit être en priorité connecté avec d'autres instants discriminants.

Définissons un masque comme une matrice $Pr = (Pr_{ij})_{ij}$ où

$$Pr_{ij} = PL_i \times PC_j \times exp(\nu|PED_i - PED_j|)$$

Le terme $exp(\nu|PED_i - PED_j|)$ revient à pénaliser les arêtes liant des instants discriminants à des arêtes qui ne le sont pas. Notons que la diagonale a toujours un poids nul, ce qui assure l'existence d'un alignement qui ne soit pas pénalisé. L'alignement retenu est un alignement favorisant le couplage entre instants de même nature, ceux faiblement discriminants et ceux qui le sont fortement. Le terme $PL_i \times PC_j$ donne plus d'importance à des instants discriminants dans le choix des arêtes. Ils apportent un faible poids aux instants ayant un poids faible. Ainsi, les masques favorisent les chemins passant par des arêtes liant des instants de même nature, en réduisant le poids des instants liant les instants faiblement discriminants. Rappelons la notation $\mathcal{A}$ décrivant l'ensemble des alignements entre deux séries. On note $(u_i, v_i) = \phi(i)$, pour $\phi \in \mathcal{A}$ Nous avons ainsi, pour DTW :
$DTW^{PM[S_1]}(S_1, S_2) = min_{\phi \in \mathcal{A}} \sqrt{\sum_{i=1}^{r} (Pr_{u_i,v_i} MW^{S^1} * X_{u_i} - S_{2v_i})^2}$

> **Remarque 65 : (Ecarts propres à chaque série)**
> *Comme précisé en introduction, les écarts définis ci-dessus sont propres à chaque série. Les distances construites ne sont pas générales et ne permettent pas a priori de calculer une distance entre deux séries quelconques. En effet, la notion de profil moyen découle directement de l'apprentissage de blocs discriminants propres à chaque série. L'utilisation de blocs d'arêtes sémantiques rend cette approche impossible à mettre en œuvre.*

> **Remarque 66 : (Axiome d'identité)**
> *Un second défaut de ce type d'écart réside dans le fait qu'une série comparée à elle-même n'a pas un écart nul. Ceci contredit l'axiome d'identité d'un indice de dissimilarité. En effet, une série est comparée à son profil moyen. Si le profil moyen s'écarte de la série initiale, alors l'écart est non nul.*

Pour pallier ce problème, nous pourrions soustraire à la quantité calculée l'écart entre la série et le profil moyen. Le problème qui se pose alors est que cette approche tend à favoriser les séries qui s'éloigne le plus de leur profil moyen, donc, qui présentent le plus de différences avec les propriétés de la classe. Nous allons donc définir une autre manière de définir des dissimilarités entre séries temporelles.

5.3.b Distances fondées sur les masques discriminants

Une autre manière de définir des distances entre séries temporelles consiste à utiliser les appariements appris. Ces appariements pointent les événements qui peuvent être liés. Cette mesure est fondée sur la somme des écarts pondérés entre les instants j d'une série S^2 et les instants i d'une série S^1.

Calcul d'une mesure de dissimilarité à partir d'un alignement discriminant A partir des blocs appris, il est possible de définir un alignement discriminant. Comme nous l'avons spécifié au paragraphe 5.3.a, un alignement est la donnée de deux vecteurs u et v de même taille r comprise entre n et 2n et d'une fonction $\Phi : \{1..r\} -> \{1..n\}^2$ telle que $max(\Phi(i+1) - \Phi(i)) = 1$ et $min(\Phi(i+1) - \Phi(i)) \leq 0$. Par programmation dynamique, nous recherchons au sein des blocs moyens appris, un alignement Φ maximisant la fonction coût suivante :

$$\sum_{k=1}^{r} \bar{M}_{\Phi(k)} - \lambda$$

Cet alignement vise à chercher le chemin maximisant le poids du bloc moyen, avec une contrainte de régularisation évitant des chemins trop longs. La constante $\lambda > 0$ permet de moduler l'importance relative de la longueur du chemin par rapport à l'importance de la maximisation des poids. Une valeur faible de λ pénalise fortement les poids très faibles, tandis qu'une valeur plus forte apporte plus de souplesse à leur égard. Elle vise à pénaliser les chemins longs. Si nous considérons le bloc moyen de chaque série, l'alignement obtenu est un chemin au sein de la matrice de poids associée à cette série. Ce chemin correspond à la façon d'aligner les instants des séries de la classe à ceux de la série de référence S^l. Les instants comportant du délai sont conservés.

Mesure de l'écart entre les instants de deux séries Nous pouvons calculer la matrice des écart entre les séries.

$$\Delta_{ij} = (S_i^1 - S_j^2)^2$$

Nous calculons alors la somme des écarts pondérés par un masque discriminant. Plusieurs options se présentent pour le choix des masques discriminants. Nous pouvons utiliser directement la matrice apprise des appariements $\bar{M}$ pondérée par les poids discriminants des instants $P = (p_i)_{i \in \{1..T\}}$, ou par les masques entropiques appris M_H. Le masque choisi est noté $p_i Pr_{i,j}$

On définit finalement un écart entre deux séries comme étant égal à :

$$DPr = \sqrt{\sum_{i,j=1}^{T} p_i Pr_{i,j} \Delta^{S_1,S_2}{}_{i,j}} \tag{92}$$

Définition d'une mesure de dissimilarité L'axiome d'identité n'est pas vérifié car au sein d'une série, les différentes arêtes activées apportent de la variabilité au sein de l'écart. Nous proposons la mesure de dissimilarité suivante D* fondée sur les bandes de Sakoe–Chiba. La métrique proposée consiste à considérer, pour tous les rayons r possibles, tous les écarts appartenant à une bande de Sakoe–Chiba de rayon r, et de choisir le plus petit écart. Nous notons SC^r la matrice associée à ces bandes. Son terme général ${SC^r}_{ij} = 1$ si $|i-j| \leq r$, 0 sinon.

$$D* = min_{r \in 1..n} \sqrt{\sum_{i,j=1}^{T} p_i \frac{Pr_{i,j} SC^r_{i,j}}{\sum_k Pr_{i,k} SC^r_{i,k}} \Delta_{i,j}^{S_1,S_2}} \tag{93}$$

Nous pouvons remplacer l'écriture matricielle de la bande de Sakoe–Chiba par l'écriture suivante.

$$D* = min_{r \in 1..n} \sqrt{\sum_{i=1}^{T} \sum_{\substack{j \in 1..T \\ |i-j| \leq r}} p_i \frac{Pr_{i,j}}{\sum_k Pr_{i,k}} \Delta M_{i,j}^{S_1,S_2}} \tag{94}$$

Remarque 67 : (Conditions d'existence)
Une condition d'existence de la dissimilarité D est la présence de termes non nuls sur la diagonale.*

- *Dans le cas des poids entropiques M_H, la construction de ces poids assurent des poids tous non nuls, le masque matriciel charge donc la diagonale.*
- *Dans le cas des poids issus de la moyenne de la matrice d'appariements Mmoy, la présence du bloc diagonal liant une série à elle-même dans la matrice d'appariements assure des poids diagonaux non nuls.*
- *Dans le cas d'un masque quelconque (par exemple, après seuillage d'un des deux masques précédents), la normalisation en ligne de la matrice $Pr_{i,j} SC^r_{i,j}$ donne un poids de 1 à la diagonale par passage à la limite.*

Remarque 68 : (Axiome d'identité)
L'axiome d'identité est à présent respecté. L'écart euclidien pondéré ($r = 0$) entre une série S_l et elle-même est nul. Le min est donc atteint pour r=0.

Remarque 69 : (Intérêt de cette distance)
Les distances classiques sont fondées sur la notion d'alignement. Elles suggèrent une hypothèse d'ordre au sein des événements. Il peut arriver que des événements arrivent à des instants différents entre plusieurs séries, alors même que les séries sont proches. Cependant, il semble naturel, du fait que les données ont une dimension temporelle, de privilégier en priorité les séries se ressemblant sur la base des instants proches. La distance que nous définissons est en mesure d'aligner les instants selon les arêtes les plus discriminantes, à partir des poids appris, quelle que soit la position de ses instants. Cependant, elle privilégie les séries proches sur des instants peu éloignés, à l'instar des métriques classiques, telles que la DTW.

5.3.c Classification des plus proches voisins sur la base de cette distance

Nous avons extrait des deux jeux de données UMD et BME dix échantillons d'apprentissage et de Test. La table 2 donne les taux de classification associés. Chaque échantillon d'apprentissage contient 12 séries, les échantillons test contiennent 48 séries. Nous comparons les taux de classifications obtenus par la méthode 1NN appliquée à la distance euclidienne et la DTW.

TABLE 2 – Taux d'erreur de la classification k-NN

	k	D*	DE	DTW
BME	1	**0.032**	0.165	0.130
	3	0.034	0.208	0.132
	5	0.062	0.234	0.136
	7	0.079	0.297	0.191
UMD	1	**0.055**	0.173	0.121
	3	0.111	0.333	0.177
	5	0.173	0.343	0.225
	7	0.222	0.378	0.274

Nous remarquons, à travers les taux d'erreur de classification, l'efficacité des apprentissages appris, à l'aide de la métrique D, à discriminer des séries temporelles complexes. Les taux de classification correcte sont bien meilleurs pour notre approche.

6 Conclusion

Nous avons, dans ce chapitre, présenté plusieurs méthodes pour déduire des appariements appris un système de poids permettant d'évaluer le caractère discriminant de chaque instant des séries. Une première méthode fondée sur la variance, attribue à un instant un poids lié à la stabilité de son voisinage, tandis qu'une deuxième approche, fondée sur l'entropie, attribue un poids lié à la dispersion du voisinage. Dans le premier cas, les voisinages les plus homogènes sont favorisés, tandis que dans le second cas, ce sont les voisinages les plus resserrés qui sont favorisés. A partir des matrices d'appariement et de ces poids discriminants, nous avons défini de nouvelles distances fondées sur la discrimination des séries temporelles. Ces distances ont été évaluées sur la base de jeux de données synthétiques. Dans la suite de cette partie, nous allons tester une de ces nouvelles distances sur un jeu de données réelles et les confronter à deux problématiques très importantes, la prédiction précoce des classes d'appartenance et l'exploration des séries à travers la recherche des points discriminants des séries.

Chapitre 6

Applications à des données de consommation électrique

Nous avons défini, dans le chapitre précédent, une nouvelle métrique $D*$. Nous proposons dans ce chapitre d'appliquer cette métrique sur un jeu de données réelles issues de relevés quotidiens de consommation électrique sur une période d'un an, où les séries présentent des profils globaux très dissimilaires au sein des classes et faiblement différentiables entre les classes une très forte variabilité. Nous souhaitons faire de la classification supervisée à partir de ces séries temporelles. Nous montrons l'efficacité de notre métrique pour la prédiction précoce d'un pic de consommation et la discrimination de séries en fonction de la saison.

Dans le chapitre précédent, nous avons exploré une méthode de classification fondée sur les appariements appris et sur un système de poids discriminants qui en découlait. Nous avons observé les très bons résultats de cette méthode pour la classification de séries temporelles extraites de jeux de données simulés. Nous allons dans ce chapitre tester la distance apprise pour la classification de séries temporelles extraites d'un jeu de données réelles. Avec l'épuisement actuel des ressources énergétiques, et l'augmentation des tarifs de l'énergie qui en découle, le problème de l'optimisation de la consommation électrique, passant par un suivi plus personnalisé et plus spécialisé n'a jamais été autant d'actualité. En particulier, l'objectif est de mieux comprendre la consommation des clients, et de prédire à l'avance un besoin plus important en énergie. Nous nous sommes intéressés, dans le cadre de la thèse, à des données issues de relevés quotidiens de la consommation électrique d'un foyer. Les relevés sont effectués à intervalles réguliers, et couvrent une période d'un an. Dans ce jeu, les séries temporelles sont des relevés de consommation électrique. Chaque instant correspond à une tranche de 10 minutes durant laquelle la consommation est relevée et une série correspond au relevé d'une journée. Nous nous intéressons à deux types de problématiques :

- Dans le but d'optimiser la production électrique, les fournisseurs souhaiteraient pouvoir anticiper les pics de consommation. En particulier, un pic assez important peut apparaître entre 18h et 20h. Pouvons-nous alors utiliser les appariements appris pour faire de la prédiction précoce d'un potentiel pic de consommation ?
- L'apprentissage des appariements caractéristiques, voire discriminants, nous permet-il

de mettre à jour des caractéristiques particulières au sein de certaines séries ?
Dans le cadre de ces données, nous nous sommes questionnés sur ces deux points : l'existence d'un pic de consommation en fin d'après-midi et sa prédiction précoce, d'une part, et d'autre part la détection de tendances de consommation saisonnières. Nous présentons dans ce qui suit les deux applications. Elles soulèvent deux problématiques totalement différentes, bien que fondées sur les mêmes jeux de données. Ainsi, les paramètres de la méthode d'apprentissage des alignements seront adaptés à chaque problème.

Notons que, dans ce chapitre, nous ne nous intéressons qu'à la version booléenne de l'algorithme.

1 Présentation des données

Nous introduisons dans cette section les données auxquelles nous nous sommes intéressés, et la spécificité des deux types de problèmes.

1.1 Généralités

Le jeu de données étudié correspond à un relevé quotidien de la consommation électrique d'un foyer sur presque toute l'année 2007 (349 jours), avec un échantillonnage consistant en un relevé toutes les 10 minutes. Le relevé d'une journée a une structure de séries temporelles. Le jeu se caractérise donc par un ensemble de 349 séries temporelles, chacune décrite par 144 instants. La date du relevé pour chacune des séries est connue. Les séries ne sont pas regroupées en classe. Cependant, du fait de notre connaissance des dates des relevés, nous pouvons créer arbitrairement des classes de sorte à extraire certaines informations. Ces séries ont la particularité de présenter beaucoup de variabilité. Par exemple, lors d'une journée, nous pouvons observer une forte consommation à un instant donné, tandis que pour d'autres journées, cette consommation n'apparaît pas. Il y a de plus beaucoup de variabilité à l'échelle des instants (heure du coucher, heure du réveil...), et l'ordre d'arrivée des événements peut être perturbé. Ce type de série n'est pas compatible avec la notion d'alignements proposée lors des approches de type DTW.

A partir de ces séries, nous avons cherché à construire deux jeux distincts correspondant aux deux tâches considérées.

1.2 Construction de deux jeux

Dans le cadre de ces données, les fournisseurs d'énergie ont pour objectif d'extraire une structure permettant de mieux comprendre la consommation pour mieux l'anticiper. A l'échelle des variations saisonnières, y a-t-il une structure particulière commune au sein des saisons ? Y a-t-il un événement qui permettrait de prédire un futur pic de consommation entre 18h et 20h ?

1.2.a Tendances saisonnières

Pour répondre au premier problème de la caractérisation des saisons, nous avons décidé, en fonction des dates du calendrier, de séparer les séries en deux saisons, une saison froide (d'octobre à avril), et une saison chaude de (mai à septembre), pour en extraire les événements

discriminants. Ce qui est d'intérêt est d'extraire un profil discriminant pour chacune des deux saisons. Nous avons à l'issue du processus d'apprentissage, effectué une classification fondée sur ces profils pour vérifier la capacité de notre approche à discriminer les séries. Ces tests de classification ont été effectués sur la base de trois tirages aléatoires d'un échantillon de 60 séries par classe pour le jeu d'apprentissage et de 30 séries pour le jeu test, par une méthode de classification de type "k plus proches voisins".

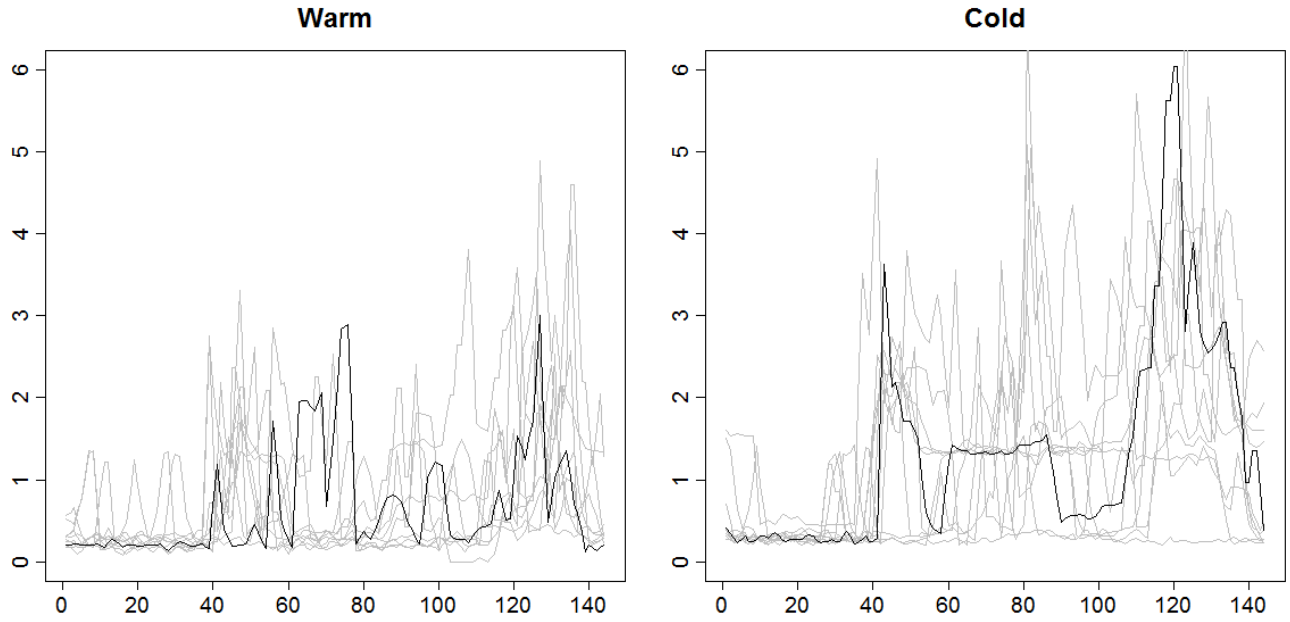


FIGURE 52 – Consommation électrique pour des séries des classes *Warm* et *Cold* du jeu de données CONSSEASON.

1.2.b Pic de consommation

Pour répondre au second problème d'anticipation d'un pic de consommation, nous avons divisé les séries en deux catégories : celles qui ont une consommation forte sur la période 18h-20h et celles qui ont une consommation faible. Pour cela, nous avons dans un premier temps, calculé la moyenne pour chaque série de ces valeurs. Nous déterminons la médiane sur cet ensemble de 349 moyennes. Enfin, en laissant un intervalle de 20% entre les deux, nous divisons la population en deux classes : les 40% de séries prenant en moyenne les valeurs les plus hautes sur cette période et les 40% prenant les valeurs les plus faibles.) Ce qui nous intéressait dans ce cas était d'être capable de prédire un éventuel pic de consommation à l'avance, c'est-à-dire de reconnaître, au sein des instants précédant la période critique, des signes discriminant les deux comportements. Nous avons à nouveau effectué une classification fondée sur les couplages appris pour vérifier la capacité de notre approche à discriminer les séries. Les tests de classification ont été effectués sur la base de trois tirages aléatoires d'un échantillon de 30 séries par classe pour le jeu d'apprentissage et de 60 séries pour le jeu test. Les séries sont classées par une méthode de classification de type "k plus proches voisins", sur la base des instants correspondant à la période 0h-16h.

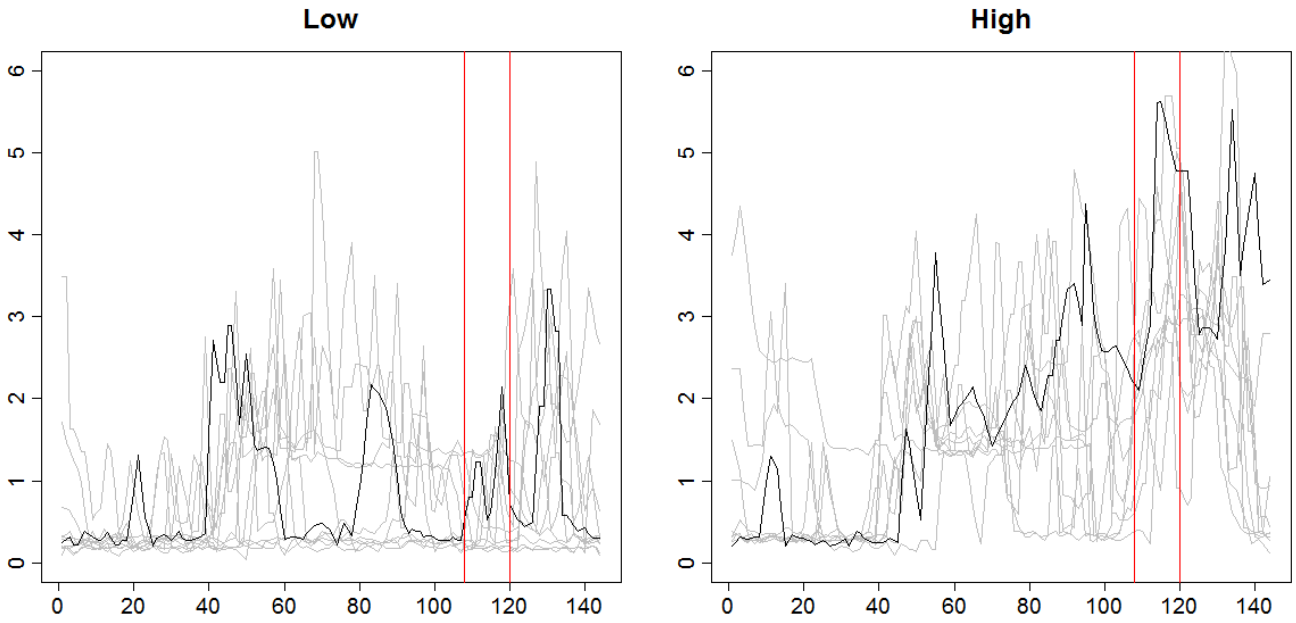


FIGURE 53 – Consommation électrique pour des séries des classes *Low* et *High* du jeu de données CONSLEVEL.

A partir de ces deux jeux de données, nous allons à présent apprendre les appariements discriminants associés par la méthode introduite précédemment.

2 Mise en place de l'apprentissage

Nous détaillons ici le choix des paramètres pour les différentes étapes de l'algorithme et les différences que nous considérons pour résoudre les deux problèmes présentés ci-dessus.

2.1 Initialisation de la matrice

Problème de la caractérisation des saisons Dans le cadre du premier problème, nous appliquons l'algorithme classique, en initialisant l'apprentissage avec un couplage complet. La sortie de l'algorithme nous donne les alignements caractéristiques au sein des classes et nous permet d'initialiser les apprentissages inter-classes.

Problème de la prédiction précoce Dans le cadre du second problème, nous appliquons l'algorithme en initialisant la matrice de voisinage avec un bloc spécifique. Afin d'apprendre des alignements dans le but de faire de la prédiction précoce, nous assignons des poids nuls à toutes les arêtes liant des instants de la seconde série situés après l'instant 16h. En effet, l'objectif étant de faire de la classification par un procédé "k plus proches voisins", nous connaissons l'information sur les séries d'apprentissage, tandis que l'information est cachée sur la base de série Test. La sortie de l'algorithme nous donne les liens discriminant les deux profils sur la base des premiers instants.

2.2 Choix du terme de tolérance

Problème de la caractérisation des saisons Dans le cadre du problème de caractérisation, nous recherchons un profil global, obtenu en faisant la moyenne des blocs associés à une série, et après application de la fonction seuil. Ainsi, la fonction seuil fait déjà office de seuil de tolérance. Nous fixons, dans ce contexte, le seuil α à 0.

Problème de la prédiction précoce Dans le cadre de la prédiction précoce, nous fixons à présent le seuil de tolérance α a une valeur plus élevée. L'objectif est d'apprendre des alignements entre paires de séries. Il faut alors éviter le sur-apprentissage. La valeur fixée pour α est de 0.05%.

2.3 Affectation à la classe des K plus proches voisins

Du fait des grandes dissimilarités entre les séries d'une même classe, qui augmente la complexité du jeu, il est judicieux de s'autoriser plus de voisins. Nous testons donc plusieurs valeurs pour le nombre de voisins K.

Pour se rendre compte de la complexité du jeu, nous représentons le résultat d'un MDS (Multidimensional-Scaling [Cox]) pour les deux distances usuelles Euclidienne et DTW.

Nous observons sur ces figures 54 et 55 un fort recouvrement des classes, qui dénote des ensembles complexes de séries temporelles. Les distances usuelles d_E et DTW sont incapables de séparer les classes efficacement.

Nous appliquons donc une classification KNN sur la base de la distance proposée au chapitre précédent. Nous présentons à présent les résultats obtenus pour les deux jeux de données.

3 Résultats

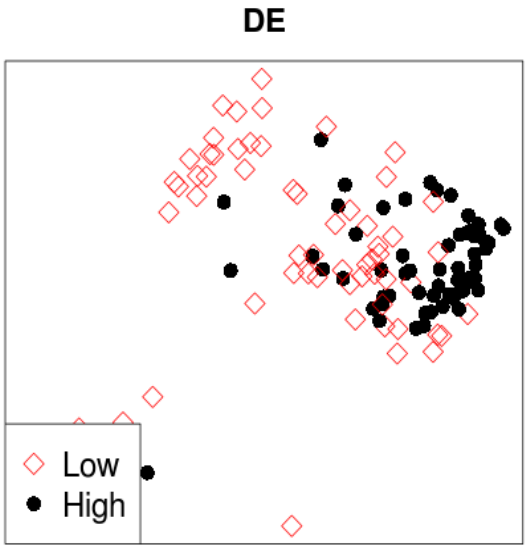
Problème de la catégorisation Nous voyons sur la figure 56, les appariements intra et inter appris au cours de l'algorithme. L'appariement intra fait état d'une figure en damier, montrant l'alternance entre des zones de faibles et fortes consommation. L'appariement inter a dégagé l'importance de la zone marquée en rouge pour la discrimination de ces séries.

Nous observons sur la figure 57, que les deux classes sont bien séparées. Cela met en avant la capacité de la métrique à caractériser les classes.

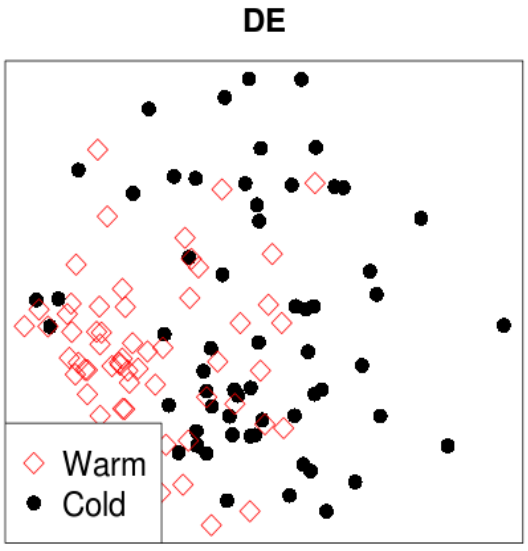
TABLE 3 – Taux d'erreur de la classification K plus proches voisins (%)

k	DE	DTW	D
1NN	23.9	28.3	**9.4**
3NN	22.8	31.1	12.8
5NN	20.0	30.0	20.5
7NN	22.2	30.6	11.1

Les résultats de classification étayent ce propos. La métrique apprise surpasse largement les métriques usuelles dans des tâches de classification.

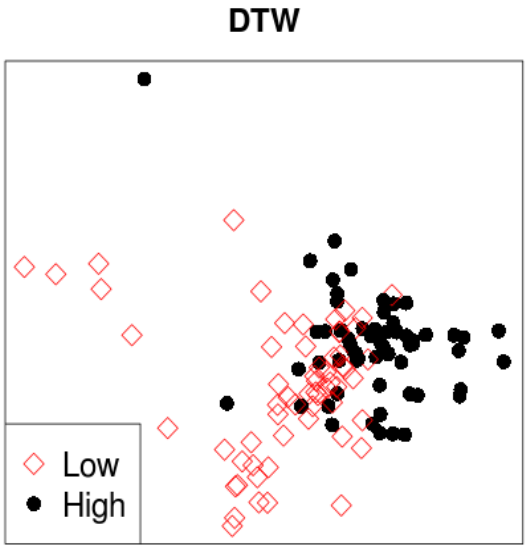


(a) Jeu Level

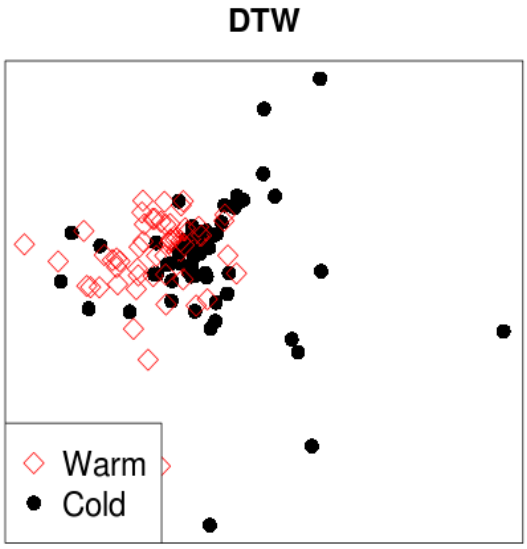


(b) Jeu Saison

FIGURE 54 – Les proximités entre les séries temporelles induites par la DE pour les deux jeux CONSLEVEL et CONSSEASON.

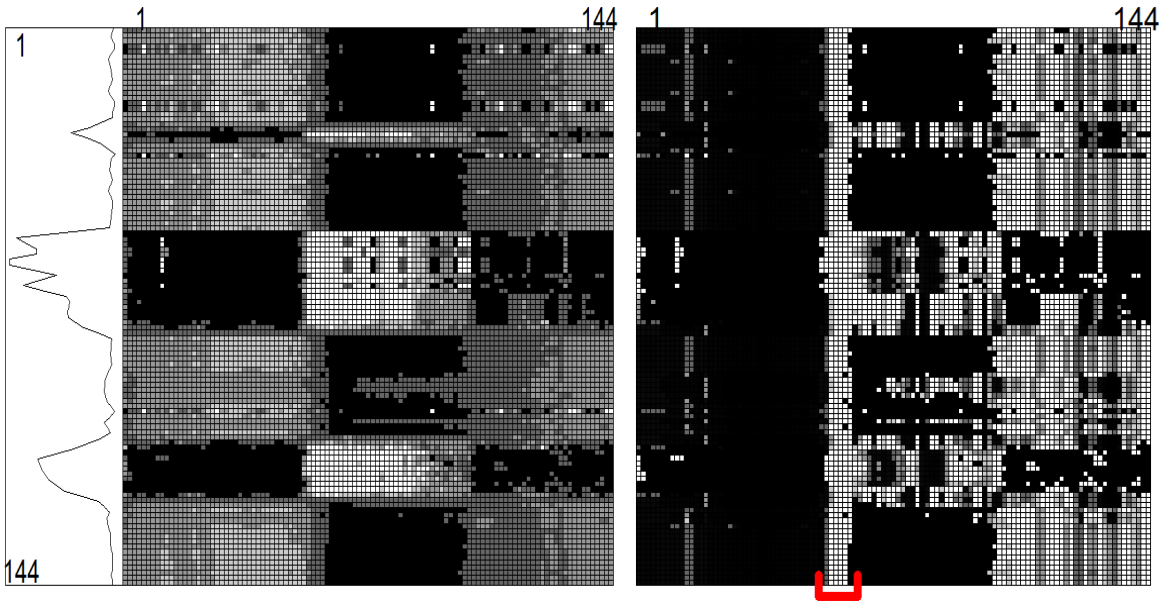


(a) Jeu Level



(b) Jeu Saison

FIGURE 55 – Les proximités entre les séries temporelles induites par la DTWpour les deux jeux CONSLEVEL et CONSSEASON.



(a) Appariement intra (b) Appariement inter

FIGURE 56 – Appariements appris pour une série "faible consommation" à l'issue d'une itération

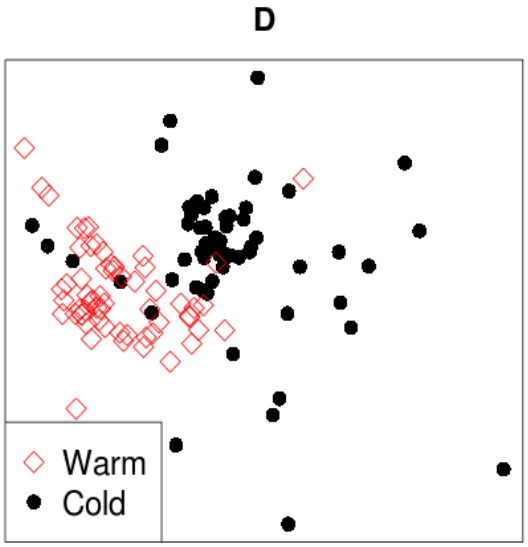


FIGURE 57 – Les proximités induites par la métrique apprise (tendances saisonnières).

3.1 Problème de la prédiction précoce

TABLE 4 – Taux d'erreur de la classification K plus proches voisins (%)

k	DE	DTW	D
1NN	30.6	28.9	5.6
3NN	26.7	26.1	4.4
5NN	23.3	23.9	2.8
7NN	23.3	23.3	**1.7**

Les résultats de classification pour la tâche de prédiction précoce montrent à nouveau une nette amélioration des taux de séries mal classées. Nous observons une amélioration des taux de classification d'un facteur 10.

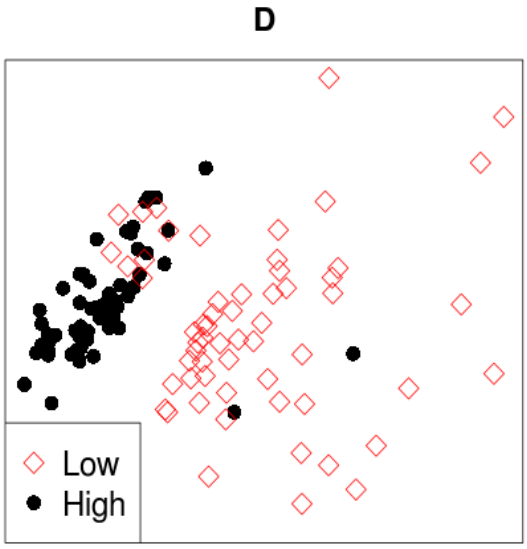


FIGURE 58 – Les proximités induites par la métrique apprise (prédiction précoce)

Nous observons à nouveau sur la figure 57, une très nette séparation des classes, qui met en avant la capacité de la métrique à prédire un épisode de forte consommation au sein des séries.

4 Conclusion

Les résultats montrent que la méthode d'apprentissage des appariements et la distance dérivée sont très performantes pour les tâches d'exploration et de prédiction précoce, tant sur des jeux simulés que sur des séries temporelles réelles de structure complexe. Les taux de classification observés sont drastiquement supérieurs aux résultats obtenus pour les distances classiquement utilisées à l'instar de la distance euclidienne et la DTW. Les masques

qui ressortent de l'apprentissage des appariements discriminants permettent de dégager des éléments pour l'interprétation des liens entre et au sein des classes.

Conclusion de la partie III

Nous avons proposé dans cette partie une nouvelle métrique discriminante. Dans le cadre d'une classification de type k-NN, cette métrique s'est avérée être très efficace pour la classification de séries temporelles. Dans le cadre d'un jeu de données complexes, où les séries peuvent présenter des profils proches dans les différentes classes et varier fortement au sein des classes, les méthodes usuelles échouent pour la classification de telles séries. Notre métrique donne en revanche de très bons taux de classification.

Conclusions et perspectives

Bilan

Nous nous sommes intéressés dans le cadre de cette thèse, à la structure de voisinage associée à un ensemble de séries temporelles regroupées en classes. La structure en classe induit deux types de liens entre les instants des séries temporelles, des liens structurels liés au découpage en classes, et des liens temporels liés à la succession des instants.

La notion de liens ainsi définie relie l'étude des séries temporelles aux travaux sur les données contiguës. Les indices usuels d'autocorrélation spatiale, à savoir les indices de Moran et de Geary, et les généralisations de la variance liées à ces indices sont étudiés sous l'angle des séries temporelles.

Ce travail de thèse a mis en évidence l'importance de la notion d'appariements, en vue de l'exploration de classes de séries temporelles. L'utilisation de voisinages définis a priori induit des hypothèses très fortes sur la structure des données, à savoir un respect de la chronologie et une difficulté à distinguer les liens temporels des liens structurels. Le fait d'introduire la notion de forme au sein de l'analyse revient à un choix particulier de voisinage et nous ramène à nouveau au problème de la définition d'un voisinage adapté à une tâche de classification. Afin de résoudre ce problème, il a été mis en évidence l'aspect essentiel qu'est l'apprentissage des appariements.

La seconde partie de ce travail a abouti à la définition d'une méthode d'apprentissage d'appariements discriminants. La méthode d'apprentissage que nous avons proposée est originale et fondée sur un principe analogue à celui de la méthode du gradient projeté. Cependant, la différence fondamentale avec cette dernière approche réside dans l'enchaînement des étapes. Le gradient projeté recherche la direction optimale et modifie la quantité à optimiser en amont de la projection, tandis que nous calculons l'effet de la projection en amont de la modification. Le travail d'analyse des différentes variantes a conduit à la définition de deux types d'approches itératives, une approche booléenne et une approche progressive. Un autre aspect novateur de notre méthode réside dans l'enchaînement d'un apprentissage des arêtes caractéristiques au sein des classes, et d'un apprentissage des arêtes différentielles entre les classes.

La convergence de l'approche booléenne a été facilement vérifiée ; cependant, dans le cadre de l'approche progressive, nous avons restreint la preuve au cadre d'une unique arête modifiée. Le cadre général a été vérifié expérimentalement. En termes de complexité, nos approches sont équivalentes aux approches du gradient projeté, qui est une référence pour ce type de problèmes d'optimisation.

Les appariements ainsi construits ont permis la définition de plusieurs systèmes de poids associés à tous les instants des séries, qui évaluent le caractère discriminant des instants.

Ces systèmes de poids permettent l'amélioration des distances usuelles, telles que la distance euclidienne et la DTW. Les appariements ont également été utilisés pour la définition d'une nouvelle mesure de dissimilarité.

L'efficacité de ces méthodes pour la caractérisation et la classification de séries temporelles a été vérifiée dans le cas de données réelles, issues de relevés de consommation électrique. Nous nous sommes intéressés à des tâches réputées difficiles telles que la prédiction précoce, ainsi qu'à des tâches de caractérisation et de classification visant à extraire les instants caractérisant les classes.

Pour terminer ce bilan, nous revenons sur les modèles de Markov cachés, très fréquemment utilisés pour ces problèmes de classification et de modélisation des séries temporelles. Un des problèmes fondamentaux de la classification de séries temporelles est lié aux dépendances temporelles entre accroissements et amplitudes, bien explorés par les techniques auto-régressives, auxquelles se surajoutent souvent des dépendances spatiales. Le démasquage a priori, par des techniques exploratoires, de telles structures cachées est fondamental et souvent beaucoup plus informatif, en vue d'une modélisation ultérieure, que le HMM, trop peu explicite, notamment en ce qui concerne les liens temporels et les liens spatiaux (surtout en cas de dépendances subtiles comme le renouvellement temporel ou spatial, non markovien). Les modèles HMM nécéssitent trop de connaissances a priori des séries, et nécessiteraient une modélisation approfondie en amont. L'alternative proposée permet de prendre en compte des données complexes, i.e., non liées à l'existence d'un profil similaire et différentiel des classes, sans connaissance a priori, tout en facilitant l'insertion d'une connaissance supplémentaire au moment de l'initialisation du procédé. Nous proposons dans la suite quelques améliorations possibles de la méthode, liées notamment à la complexité de la méthode décrite dans le manuscrit et à son extension au cadre de l'analyse non supervisée.

Perspectives

Accélération du processus d'apprentissage

Le système que nous proposons, à l'instar de toutes les méthodes d'apprentissage d'alignements multiples qui ne se limitent pas à un apprentissage paire à paire, est assez coûteux en termes de complexité et présente des difficultés de passage à l'échelle. Nous nous limitons en pratique à des séries relativement courtes (quelques centaines d'instants) et à un nombre restreint de séries (une centaine de séries). Les méthodes d'ancrage, telles que la méthode Chaos appliquée à l'algorithme DIALIGN (Brudno, 2004), visent à initialiser les matrices d'alignements par des diagonales de taille réduite et fixée au départ et les points d'ancrage sont étendus pour créer l'alignement final.

Dans le cadre de notre approche, nous pouvons mettre en place un cheminement analogue. Une première étape consiste à découper les séries en segments en fonction de la variation temporelle contenue dans ces segments. Une idée, proposée par Chouakria-Douzal (Chouakria-Douzal, 2003) pour résoudre des problèmes de compressions de séries temporelles, consiste à découper les segments dès que la corrélation temporelle cumulée le long du segment dépasse un certain seuil fixé au départ (voir figure 59).

A l'issue de ce découpage, nous proposons deux approches.

– La première approche consiste à initialiser tous les blocs entre deux segments par des

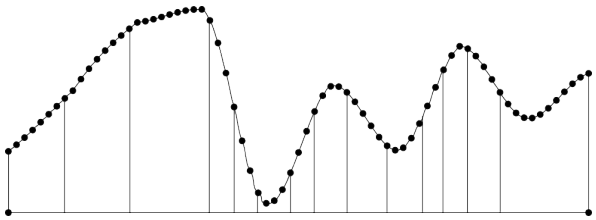
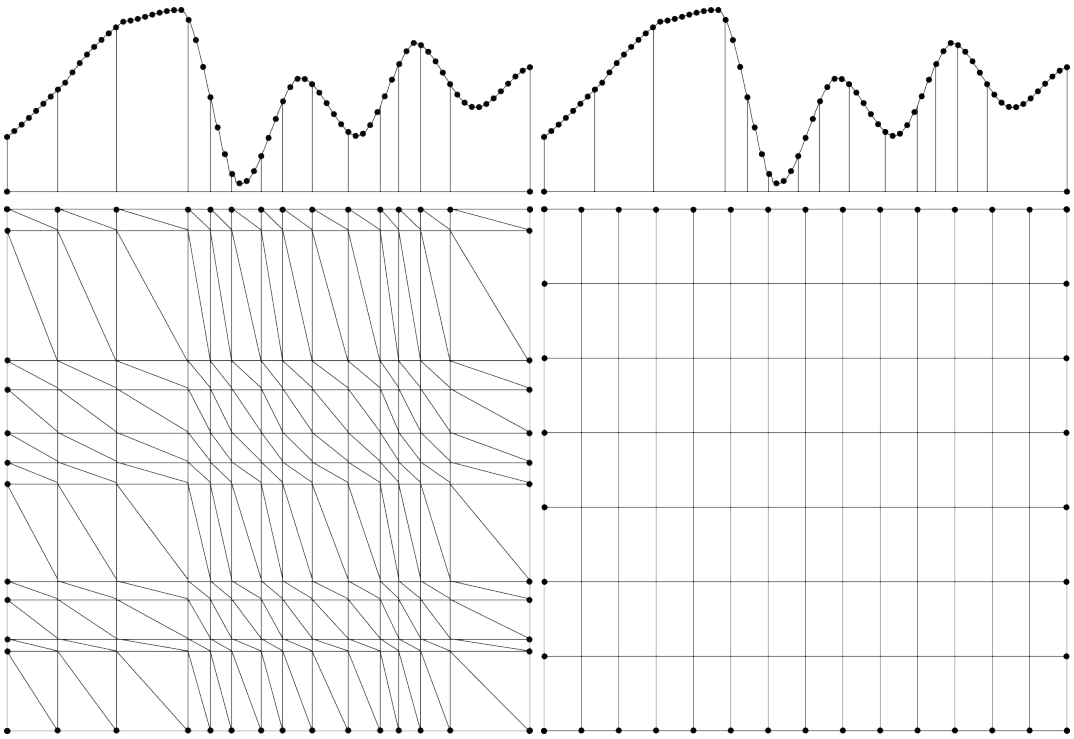


FIGURE 59 – Découpage de la série en segments

alignements diagonaux (figure 60 a), i.e. remplacer les blocs complets initiaux de la matrice par des diagonales pour chaque couple de segments. Le nombre d'arêtes associées à chaque instant est égal au nombre de segments conservés, au lieu de nT initialement. Le processus est fortement accéléré.

– La seconde approche consiste à faire une première étape d'apprentissage à partir d'une sous-matrice de dimension beaucoup plus faible (figure 60 b) (égale au nombre total de segments à la place de nT initialement), en résumant la valeur observée sur chaque segment, par sa moyenne. Les arêtes de la matrice réduite conservées au cours du processus d'apprentissage seront associées à des sous-matrices complètes au sein de la matrice de toutes les arêtes, les arêtes éliminées étant associées à des sous-matrices nulles. L'initialisation se faisant alors par des matrices creuses, à nouveau ceci accélère fortement le processus.



(a) Initialisation diagonale (b) Initialisation par réduction de la dimension

FIGURE 60 – Techniques d'accélération

Extension à l'analyse non supervisée

Une deuxième limite de notre approche est l'aspect supervisé de notre approche. En effet, la différenciation des alignements intra et inter-classes nécessite une connaissance a priori des classes. Une perspective importante de ce travail revient à étendre notre approche au cas de l'apprentissage non supervisé. La notion d'alignement est intrinsèque à l'ensemble des séries. Une solution envisagée pour ce problème consiste en l'approche suivante. A l'instar des K-moyennes, nous initialisons le processus par le choix de K séries S^l (éventuellement choisies dans l'ensemble des séries), comme dans une classification hiérarchique. Le processus se fait alors en deux étapes :

1. Un appariement intra-classe est effectué, liant toutes les séries aux centres S^l selon un couplage complet. Sur la base de la distance associée à la série S^l telle que définie dans la section 5.3.b du chapitre 5 de la partie III, toutes les séries sont affectées à la classe dont elles se rapprochent le plus du prototype.
2. Un appariement discriminant est effectué, liant toutes les séries aux centres S^l, en fonction de la classe d'affectation.
3. Nous calculons alors le profil moyen sur la base des couplages intra.
4. Nous reprenons à l'étape 1, en n'utilisant plus les séries initiales, mais les profils moyens, jusqu'à convergence des classes (comme dans les nuées dynamiques).

Travail sur la définition de prototypes de classes

L'algorithme précédent a pour vocation de faire de l'apprentissage non supervisé, sur le modèle des K-means. Cependant, les étapes 2 et 3 peuvent être mises en œuvre de manière itérative, indépendamment de l'étape 1. Ainsi, l'algorithme peut également être utilisé lorsque les classes sont connues, pour rechercher un centre de classe.

Prise en compte de la forme

Dans le travail d'apprentissage, nous nous sommes limités aux valeurs des séries de la classe. Cependant, comme nous l'avons précisé dans le chapitre 2 de la partie I, la notion d'évolution est essentielle dans le cadre des séries temporelles. Il pourrait être judicieux d'introduire les accroissements dans le processus d'apprentissage. Une solution pourrait consister à ajouter dans les données pour chaque variable un vecteur vitesse associé. L'apprentissage des couplages chercherait alors des liens minimisant la variance tant des variables que des accroissements qui en dérivent.

Approches factorielles fondées sur les appariements appris

Enfin, une perspective importante consiste à appliquer les méthodes factorielles fondées sur les structures de voisinage, et présentées au chapitre 2 de la partie I section E au cas des appariements appris, en vue de la recherche d'axes factoriels discriminants adaptés à la structure apprise pour les séries.

Annexe A

Réécriture de la covariance temporelle (Preuve)

$$CovT(X)_{jj'} = var(X^+)_{jj'} + var(X^-)_{jj'} + \frac{n-1}{n^3}x_{nj}x_{nj'} + \frac{n-1}{n^3}x_{1j}x_{1j'} - \left(cov(X^+, X^-)_{jj'} + cov(X^-, X^+)_{jj'} + \frac{1}{n^2}x_{nj}x_{1j'} + \frac{1}{n^2}x_{1j}x_{nj'}\right)$$

Preuve :

$$\begin{aligned} CovT(X)_{jj'} &= \sum_{k=1}^{n-1} \frac{1}{n}(x_{(k+1)j} - E(X_j) \\ &\quad + E(X_j) - x_{kj})(x_{(k+1)j'} - E(X'_j) + E(X'_j) - x_{kj'}) \\ &= \sum_{k=1}^{n-1} \frac{1}{n}(x_{(k+1)j} - E(X_j))(x_{(k+1)j'} - E(X_{j'})) \\ &\quad + \sum_{k=1}^{n-1} \frac{1}{n}(x_{kj} - E(X_j))(x_{kj'} - E(X_{j'})) \\ &\quad - \sum_{k=1}^{n-1} \frac{1}{n}(x_{kj} - E(X_j))(x_{(k+1)j'} - E(X_{j'})) \\ &\quad - \sum_{k=1}^{n-1} \frac{1}{n}(x_{(k+1)j} - E(X_j))(x_{kj'} - E(X_{j'})) \end{aligned}$$

Décomposons la première somme.

$$\begin{aligned}
\sum_{k=1}^{n-1} \frac{1}{n} \Big(x_{(k+1)j} - E(X_j)\Big) \Big(x_{(k+1)j'} - E(X_{j'})\Big) = & \\
& \sum_{k=1}^{n-1} \frac{1}{n} \Big(x_{(k+1)j} - E(X_j^+) \\
& + \frac{1}{n} x_{1j}\Big) \Big(x_{(k+1)j'} - E(X_{j'}^+) + \frac{1}{n} x_{1j'}\Big) \\
& = \sum_{k=1}^{n-1} \frac{1}{n} \Big(x_{(k+1)j} - E(X_j^+)\Big) \Big(x_{(k+1)j'} - E(X_{j'}^+)\Big) \\
& \quad + \sum_{k=1}^{n-1} \frac{1}{n} \Big(\frac{1}{n} x_{1j}\Big) \Big(x_{(k+1)j'} - E(X_{j'}^+)\Big) \\
& \quad + \sum_{k=1}^{n-1} \frac{1}{n} \Big(\frac{1}{n} x_{1j'}\Big) \Big(x_{(k+1)j} - E(X_j^+)\Big) \\
& \quad + \sum_{k=1}^{n-1} \frac{1}{n} \Big(\frac{1}{n} x_{1j}\Big) \Big(\frac{1}{n} x_{1j'}\Big) \\
& = var(X^+)_{jj'} + \frac{n-1}{n^3} x_{1j} x_{1j'}
\end{aligned}$$

$$\text{car } \sum_{k=1}^{n-1} \frac{1}{n^2} x_{1j} \Big(x_{(k+1)j'} - E(X_{j'}^+)\Big)$$

$$= \frac{1}{n} x_{1j} \sum_{k=1}^{n-1} \frac{1}{n} \Big(x_{(k+1)j'} - E(X_{j'}^+)\Big) = 0$$

De la même façon, nous trouvons

$$\sum_{k=1}^{n-1} m_k \left(x_{kj} - E(X_j)\right) \left(x_{kj'} - E(X_{j'})\right) = var(X^-)_{jj'} + \frac{n-1}{n^3} x_{nj} x_{nj'}$$

Enfin, les deux dernières sommes se transforment également

$$\begin{aligned}
(x_{kj} - E(X_j)) \qquad (x_{(k+1)j'} - E(X_{j'})) &= (x_{kj} - E(X_j^-) + \frac{1}{n} x_{nj})(x_{(k+1)j'} - E(X_{j'})) \\
&= (x_{kj} - E(X_j^-))(x_{(k+1)j'} - E(X_{j'})) + \frac{1}{n} x_{nj}(x_{(k+1)j'} - E(X_{j'})) \\
&= (x_{kj} - E(X_j^-))(x_{(k+1)j'} - E(X_{j'}^+) + \frac{1}{n} x_{1j}) \\
&\quad + \frac{1}{n} x_{nj}(x_{(k+1)j'} - E(X_{j'})^+) + \frac{1}{n^2} x_{nj} x_{1j'}
\end{aligned}$$

or $\sum_{k=1}^{n-1} \frac{1}{n}(x_{(k+1)j'} - E(X_{j'}^+)) = \sum_{k=1}^{n-1} \frac{1}{n}(x_{(k)j'} - E(X_{j'}^-)) = 0$. Nous avons donc

$$\sum_{k=1}^{n-1} \frac{1}{n}(x_{kj} - E(X_j))(x_{(k+1)j'} - E(X_{j'})) = cov(X^+, X^-)_{jj'} + \frac{1}{n^2} x_{nj} x_{1j'}$$

Ainsi

$$CovT(X)_{jj'} = var(X^+)_{jj'} + var(X^-)_{jj'} + \frac{n-1}{n^3} x_{nj} x_{nj'} + \frac{n-1}{n^3} x_{1j} x_{1j'} - \left(cov(X^+, X^-)_{jj'} + cov(X^-, X^+)_{jj'} + \frac{1}{n^2} x_{nj} x_{1j'} + \frac{1}{n^2} x_{1j} x_{nj'} \right)$$

□

En particulier, pour une valeur de n assez grande,

$$CovT(X) \cong var(X^+) + var(X^-) - cov(X^+, X^-) - cov(X^-, X^+)$$

Toujours dans l'hypothèse où les poids sont égaux, ainsi que les différences de temps entre deux instants, revenons aux matrices de variance covariance étudiées dans le cadre de l'ACPVI.

$$Y_{ij} = (x_{ij} - E(X_j)) \tag{95}$$

$$Y^+_{ij} = \frac{1}{\delta}(x_{i+1j} - E(X^+_j)) \tag{96}$$

$$Y^t D Y^+_{jj'} = \frac{1}{n\delta} \sum_{k=1}^{n-1} (x_{kj} - E(X_j))(x_{(k+1)j'} - E(X^+_{j'})) \tag{97}$$

Or nous avons remarqué que :

$$\sum_k (x_{kj} - E(X_j))(x_{(k+1)j'} - E(X^+_{j'})) = cov(X^+, X^-)_{jj'} \tag{98}$$

$$\text{De plus, } \sum_k (x_{kj} - E(X_j))(x_{(k)j'} - E(X^-_{j'})) = cov(X^-)_{jj'} \tag{99}$$

Nous remarquons que $Y^t D Y^-$ est la matrice de variance-covariance de X^-, et $Y^t D Y^+$ est la matrice de covariance entre X^+ et X^-. Nous avons donc $Var(V)$=$Var(X^+) + Var(X^-)$ - $(Cov(X^+, X^-) + Cov(X^-, X^+)) \cong CovT(X)$
De plus, $Var_V(X) = Var(X^-)$ - $Cov(X^+, X^-)$

Annexe B

Evolution de la variance en fonction de la pénalisation

1 Formule explicite de l'évolution de la variance

On fait dans la suite l'hypothèse que les données sont univariées, la démonstration est identique dans le cas multivarié. Rappelons quelques notations. On note x_i la ième valeur observée, et $\bar{x}^i$ la moyenne pondérée des valeurs observées sur le voisinage de x_i. Notons $p_{kk'}$ le poids de l'arête kk', quand les poids de voisinage sont normalisés en ligne, $\tilde{p_{kk'}}$ le nouveau poids obtenu après pénalisation de l'arête ii' et renormalisation en ligne. β est le taux de pénalisation de l'arête.

$\Delta_{ii'}(\beta)$ évalue l'évolution de la variance en fonction des modifications. La question que nous nous posons est de connaître le signe de la fonction Δ sur l'intervalle [0,1].

Etant donné que dans ce calcul, un seul poids est pénalisé avant une renormalisation en ligne, seuls les poids sur la ligne i vont être modifiés. On notera βP le poids $\tilde{p_{ii'}}$ associé à l'arête ii' après pénalisation et avant d'effectuer la normalisation. En notant $\beta_{rr'}$ les facteurs d'évolution des poids (i.e. tels que $\tilde{p_{rr'}} = \beta_{rr'} p_{rr'}$), on a

$$\begin{aligned} \beta_{ii'} &= \frac{\beta}{1-(1-\beta)P} \\ \text{si } r' \neq i' \ \beta_{ir'} &= \frac{1}{1-(1-\beta)P} \\ \text{si } r \neq i \ \beta_{rr'} &= 1 \end{aligned}$$

On rappelle l'expression de la variance V et de la variance tronquée $V_{ii'}(\beta)$ après pénalisation de l'arête ii' par un facteur β.

$$\begin{aligned} V = \sum_r p_r (x_r - \bar{x}^r)^2 \text{ avec } \bar{x}^r &= \sum_{r'} p_{rr'} x_{r'} \\ V_{ii'}(\beta) = \sum_r p_r (x_r - \tilde{\bar{x}})^2 \text{ avec } \tilde{\bar{x}} &= \sum_{r'} \tilde{p_{rr'}} x_{r'} \\ \Delta_{ii'}(\beta) &= V - V_{ii'}(\beta) \end{aligned}$$

Ainsi,

$$\begin{aligned}\Delta_{ii'} &= p_i(x_i-\bar{x}^i)^2 - p_i(x_i-\tilde{\bar{x}}^i)^2\\ &= \Big(x_i-\sum_k p_{ik}x_k\Big)^2-\Big(x_i-\sum_k p_{ik}\beta_k x_k\Big)^2\\ &= \Big(\sum_k p_{ik}(\beta_k-1)x_k\Big)\Big(2x_i-\sum_k p_{ik}(\beta_k+1)x_k\Big)\end{aligned}$$

On remarque que $\Delta_{ii'}(1)=0$ et $\Delta_{ij}(0)=WC_{ij}$

$$\begin{aligned}\sum_k p_{ik}(\beta_k-1)x_k &= \frac{(1-\beta)p_{ij}}{1-(1-\beta)p_{ij}}\bar{x}^i-\Big(\frac{(1-\beta)p_{ij}}{1-(1-\beta)p_{ij}}+(1-\beta)\Big)p_{ij}x_j\\ &= \frac{(1-\beta)p_{ij}}{1-(1-\beta)p_{ij}}\Big(\bar{x}^i-(1+\beta p_{ij})x_j\Big)\end{aligned}$$

$$\sum_k p_{ik}(\beta_k+1)x_k = \frac{2-(1-\beta)p_{ij}}{1-(1-\beta)p_{ij}}\bar{x}^i+\frac{\beta-1}{1-(1-\beta)p_{ij}}(1+(\beta+2)p_{ij})p_{ij}x_j$$

D'où

$$\begin{aligned}\Delta_{ij}(\beta) &= \frac{(1-\beta)p_{ij}}{(1-(1-\beta)p_{ij})^2}\Big(\bar{x}^i-(1+\beta p_{ij})x_j\Big)\\ &\Big(2x_i(1-(1-\beta)p_{ij})-(2-(1-\beta)p_{ij})\bar{x}^i-(\beta-1)(1+(\beta+2)p_{ij})p_{ij}x_j\Big)\end{aligned}$$

En particulier, cette expression est du même signe que l'expression

$$\begin{aligned}&(1-\beta)(\bar{x}^i-x_j-\beta p_{ij}x_j)\\ &(2(1-(1-\beta)p_{ij})x_i-(2-(1-\beta)p_{ij})\bar{x}^i-(\beta-1)(1+(\beta+2)p_{ij})p_{ij}x_j)\end{aligned}$$

qui est un polynôme de degré 4, dont 1 et $\frac{\bar{x}^i-x_j}{p_{ij}x_j}$ sont deux racines évidentes.

On observe sur un exemple les valeurs prises par la fonction $\Delta ij(\beta)$ avec x_i=-1, x_j=10, $\bar{x}^i$=-12, p_{ij}=10^{-3}, proche du poids initial.

2 Etude du signe de la fonction de pénalisation

$$f:\beta\mapsto \qquad V_{m^{ll'}_{ii'}}-V(\beta\, m^{ll'}_{ii'}) \tag{100}$$

où $\beta\in[0,1]$ et $V(\beta\ m^{ll'}_{ii'})$ est la variance totale obtenue après pénalisation par un facteur β du lien (i,i',l,l') et renormalisation des liens $(i,i'',l,l')(i''=1,\cdots,T)$ pour satisfaire aux contraintes. f vérifie que $f(1)=0$ et $f(0)=C^{ll'}_{ii'}$.

Nous introduisons deux autres fonctions δ_1 et δ_2 définies pour le triplet (i,j,l) par :

$$\begin{aligned}\delta_1(x^{l'}_{i'j}) &= x^{l'}_{i'j}-\sum_{r=1}^{M}\sum_{t=1}^{T}m^{lr}_{it}x^r_{tj}\\ \delta_2(i',l') &= \frac{m^{ll'}_{ii'}}{2(1-m^{ll'}_{ii'})}(1+\frac{2m^{ll'}_{ii'}x^{l'}_{i'j}}{\delta_1(x^l_{ij},x^{l'}_{i'j})})\end{aligned}$$

La propriété suivante donne des conditions sous lesquelles le signe de la dérivée évaluée en 1 $f'(1)$ est différent du signe de la fonction en 0 $f(0)$.

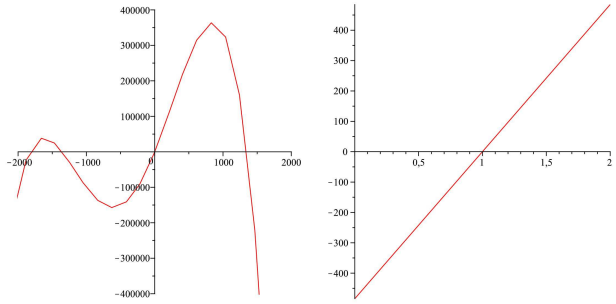


FIGURE 61 – Valeurs prises par le polynôme

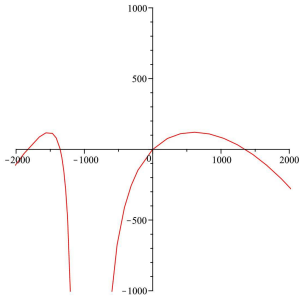


FIGURE 62 – Valeurs de $\Delta_{ij}(\beta)$ pour β quelconque

Proposition 70 :
Soit Λ *le produit* $\delta_1(x_{ij}^l) \times delta_1(x_{i'j}^{l'})$. *Alors :*

1. $\text{signe}(-f'(1)) \neq \text{signe}(\Lambda) \Leftrightarrow 0 < \delta_1(x_{i'j}^{l'}) < m_{ii'}^{ll'} x_{i'j}^{l'}$
2. $\text{signe}(f(0)) \neq \text{signe}(\Lambda) \Leftrightarrow 0 < \delta_1(x_{ij}^l) < \delta_2(i', l')$ ou
   $0 < -\delta_1(x_{ij}^l) < -\delta_2(i', l')$

Preuve. Ces cas extrêmes se rencontrent très rarement, nous préciserons également dans quelles conditions on peut les rencontrer.
La preuve est fondée sur une fonction polynomiale g dont les variations sont similaires à f.

$$g(\beta) = (1-\beta)(\delta_1(x_{i'j}^{l'}) - \beta m_{ii'}^{ll'} x_{i'j}^{l'})((2-(1-\beta))\delta_1(x_{ij}^l) \\ -(1-\beta)m_{ii'}^{ll'}(x_{ij}^l + x_{i'j}^{l'} + (\beta+2)m_{ii'}^{ll'} x_{i'j}^{l'}))$$

f et g ont le même signe, ainsi que leur dérivée en 1. $\text{signe}(f'(1)) = \text{signe}(g'(1))$. La comparaison du signe de $g(0)$ et de $g'(1)$ avec le signe de Λ conclut la démonstration. ■

Pour tous les liens de $\mathcal{E}$, $f(0)$ est positif et la variance décroît pour des faibles pénalisations ($f'(1) < 0$) dès que $f(0)$ et $-f'(1)$ ont le même signe que Λ. La propriété 50 donne les conditions sous lesquelles cette égalité est vérifiée.

- Le premier cas arrive lorsque $x_{i'j}^{l'}$ est très proche des voisins de x_{ij}^l, dès que la renormalisation des poids à l'issue de la pénalisation entraîne plus d'effets que la pénalisation elle-même ; dans ces conditions, la variance peut augmenter.
- Le second cas apparaît lorsque la distance entre x_{ij}^l et ses voisins est faible, de sorte que la variation du poids de $x_{i'j}^{l'}$ écarte x_{ij}^l de sa moyenne de voisinage. Ce cas est proche

du cas de convergence. Ces cas arrivent rarement ; il suffit de faire un test au sein de l'algorithme pour éviter ces cas là.

Annexe C

Description des jeux de données simulées

Les distances usuelles sont en général fondées sur la notion d'alignement. La notion d'alignement fait deux suppositions sur la nature des séries temporelles. Deux séries temporelles sont proches si elles partagent des événements communs, et si la chronologie de ces événements est conservée. Pourtant, dans certains jeux de données, cette condition n'est pas forcément vérifiée. Deux tâches peuvent être inversées. La proximité entre deux séries peut découler dans certains cas uniquement de la présence ou non d'un ou plusieurs événements particuliers, indépendamment de leur chronologie. Pour illustrer ce problème, nous proposons d'introduire deux nouveaux jeux de données dans lesquels la chronologie n'est pas importante. Leur structure est simillaire. Ce sont deux jeux univariés, où les séries temporelles sont constituées de trois régions. : une région centrale où apparaît un large plateau et deux régions situées au début et à la fin, où peut apparaître un pic. Les classes se caractérisent par la présence et le signe de ces pics. Nous trouvons dans les deux jeux plusieurs profils au sein des classes. Nous détaillons dans la suite les deux jeux.

1 Begin - Middle - End

Ce jeu est constitué de trois classes de séries temporelles de longueur T=128. Toutes les séries sont constituées d'un plateau central, les plateaux centraux pouvant être positifs ou négatifs. L'élément discriminant est la présence ou non d'un pic positif, soit en début de série, soit à la fin.

Caractéristiques du jeu Pour toutes les classes, les instants de début D et de fin F du plateau sont variables et valent en moyenne 40 et 90. L'orientation (vers le haut ou vers le bas) fluctue au sein même des classes. Selon l'orientation du plateau, sa hauteur fluctue ou bien autour de 1 ou bien autour de -1. L'étendue des fluctuations varie en fonction des classes, avec un écart-type plus faible au sein de la classe "Begin", qu'au sein des deux autres, ce qui brise la symétrie entre les classes "Begin" et "End". Lors d'une classification, cela permet de distinguer les erreurs de classification dues à la variabilité de la bosse centrale. Nous précisons maintenant pour chaque classe la construction du pic. A l'issue de cette construction, les instants des séries sont tous perturbés à partir d'un bruit Gaussien $\epsilon(t) = \mathfrak{N}(0, 0.01)$, centré

et d'écart-type 0.01. Les séries ainsi construites sont finalement lissées par un filtre moyenneur.

Détail des classes • La classe "Middle" est caractérisée par l'absence de pic aux deux extrémités de la série. De plus, le plateau central est toujours orienté vers le haut. Le plateau est donc caractéristique de la classe. En revanche, il n'est pas discriminant, dans la mesure où environ la moitié des séries des autres classes partagent ce plateau. Ce qui discrimine la classe "Middle" vis-à-vis des autres classes est l'absence simultanée des plateaux en début et en fin de série. Il y a pour cette série un unique profil.

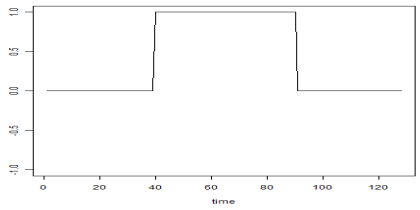


FIGURE 63 – Profil des séries de la classe "Middle".

Les instants de début D et de fin F du plateau sont variables et valent en moyenne 40 et 90. Plus précisément, D suit une loi binomiale $\mathfrak{B}(60, \frac{2}{3})$ et F est défini comme $D + 10 + \Delta$ où Δ suit une loi de Poisson $\mathfrak{P}(40)$. La hauteur du plateau est variable, et suit une loi Gaussienne $\epsilon(t) + \mathfrak{N}(1, 0.2)$ centrée en 1 et d'écart-type ≈ 0.2. Les hauteurs des plateaux sont assez fluctuantes. La figure 1 donne l'allure d'un échantillon de séries de la classe.

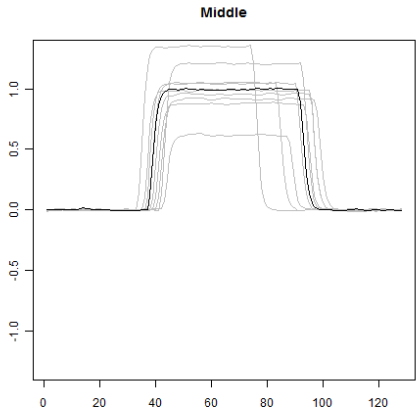


FIGURE 64 – Echantillon de séries de la classe "Middle".

• La classe "Begin" est caractérisée par un pic positif apparaissant en début de série. Il apparaît en moyenne entre les instants 113 et 118, et prend une valeur autour de 1. Le plateau central est orienté vers le haut pour certaines séries, vers le bas pour d'autres. Le pic est donc l'élément discriminant de la série. Les séries peuvent donc prendre deux types de profils.

Le pic est un petit plateau situé en moyenne entre les instants 10 et 16, ayant une valeur autour de 1. Précisément, l'instant de début du pic D suit une loi binomiale $\mathfrak{B}(15, \frac{2}{3}$ centrée

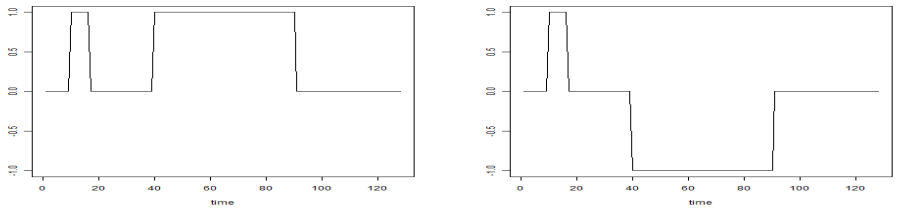


FIGURE 65 – Profils des séries de la classe "Begin".

autour de 10 et l'instant de fin F est défini comme $D+\Delta$ où Δ suit une loi de Poisson $\mathfrak{P}(4)$. La hauteur du pic suit également une Gaussienne $\epsilon(t) + \mathfrak{N}(1, 0.1)$ de moyenne 1 et d'écart-type ≈ 0.1. Le signe du plateau est équiprobable. Comme pour la classe "Middle", les instants de début D et de fin F du plateau sont variables et valent en moyenne 40 et 90. La hauteur du plateau suit une loi Gaussienne centrée soit en 1, soit en -1, selon le signe du plateau, et d'écart-type ≈ 0.05. Les hauteurs (en valeur absolue) des plateaux sont assez stables. La figure 1 donne l'allure d'un échantillon de séries de la classe.

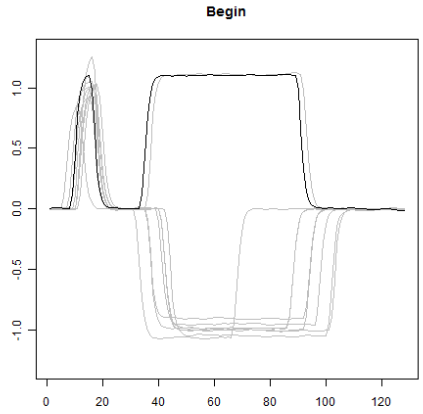


FIGURE 66 – Echantillon de séries de la classe "Begin".

• La classe "End" est caractérisée par un pic positif apparaissant en fin de série. Il est situé en moyenne entre les instants 113 et 118, et prend une valeur autour de 1. Le plateau central est orienté vers le haut pour certaines séries, vers le bas pour d'autres. Le pic est donc l'élément discriminant de la série. Les séries peuvent donc prendre deux types de profils.

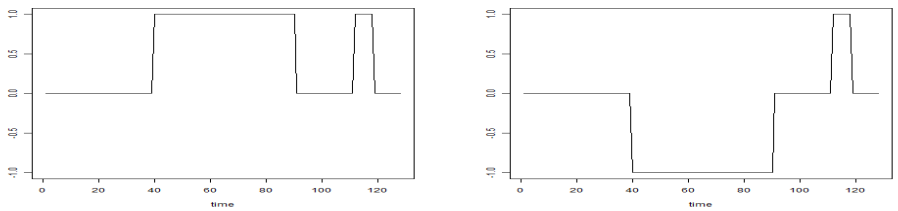


FIGURE 67 – Profils des séries de la classe "End".

Le pic est un petit plateau situé en moyenne entre les instants 113 et 118, ayant une valeur

autour de 1. Précisémment, l'instant de fin F s'exprime en fonction d'une variable binomiale $D' = \mathfrak{B}(15, \frac{2}{3}$ avec F= $128 - D'$, et D est défini comme $F - 2 - \Delta$ où Δ suit une loi de Poisson $\mathfrak{P}(4)$. La hauteur du pic suit également une Gaussienne $\epsilon(t) + \mathfrak{N}(1, 0.1)$ de moyenne 1 et d'écart-type ≈ 0.1. Le signe du plateau est équiprobable. Comme pour les autres classes, les instants de début D et de fin F du plateau sont variables et valent en moyenne 40 et 90. La hauteur du plateau suit une loi Gaussienne centrée soit en 1, soit en -1, selon le signe du plateau, et d'écart-type ≈ 0.2. Les hauteurs (en valeur absolue) des plateaux fluctuent. La figure 1 donne l'allure d'un échantillon de séries de la classe.

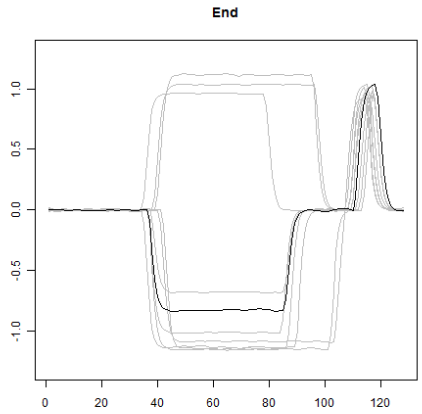


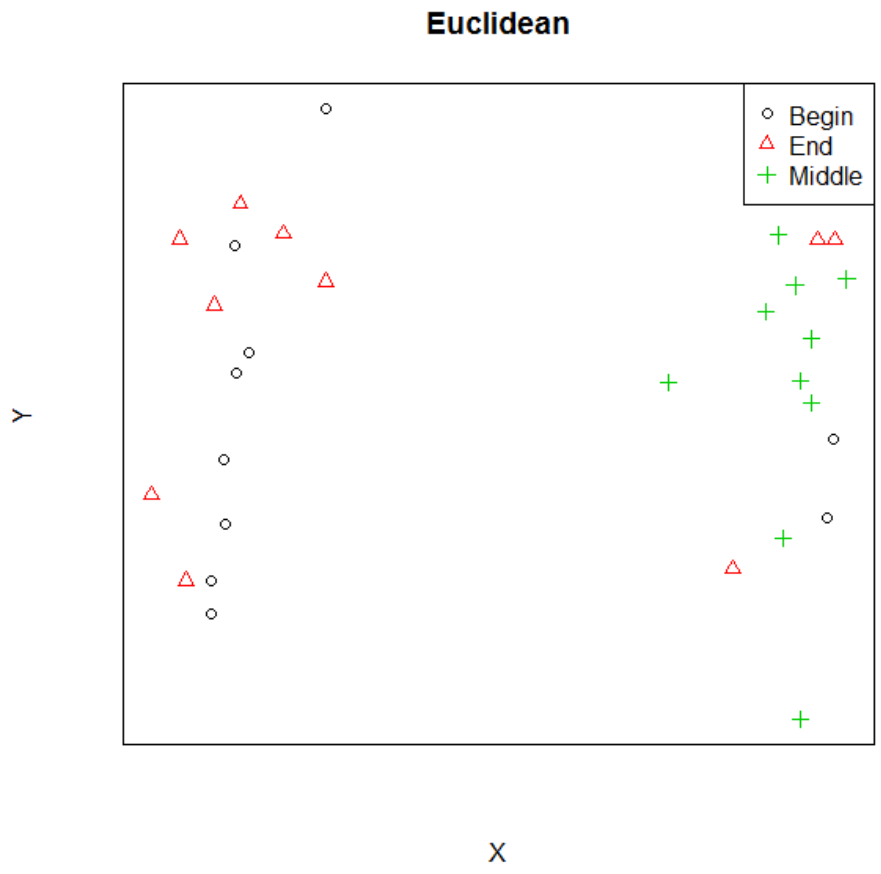
FIGURE 68 – Echantillon de séries de la classe "End".

Complexité du jeu de données La distance euclidienne, calculée entre chaque paire de série, tient compte principalement de la position de la bosse centrale. Nous voyons sur le MDS figure 1 deux groupes. Le premier correspond aux séries ayant une bosse centrale vers le haut, le second aux séries ayant une bosse vers le bas. Ce jeu est donc complexe dans le sens où la structure des classes ne s'explique pas sur la lecture des écarts au sein des instants.

2 Up - Middle - Down

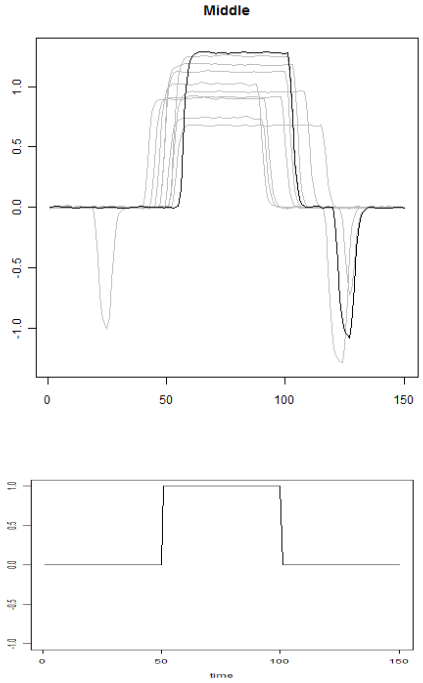
Ce jeu est constitué de trois classes de séries temporelles de longueur T=150. La structure des séries est commune aux séries du jeu précédent. Toutes les classes sont à nouveau constituées d'un plateau central pouvant être à valeurs positives ou négatives. L'élément discriminant est la présence d'un pic, soit en début de série, soit à la fin. La différence avec le jeu précédent réside dans le fait que ce qui caractérise les classes est le signe du pic et non plus sa position.

Caractéristiques du jeu Pour toutes les classes, les instants de début D et de fin F du plateau sont variables et valent en moyenne 51 et 100. L'orientation (vers le haut ou vers le bas) fluctue au sein même des classes. Selon l'orientation du plateau, sa hauteur fluctue toujours ou bien autour de 1 ou bien autour de -1. L'étendue des fluctuations est à présent constante au sein de toutes les classes. Plus précisément, D se déduit d'une loi binomiale et vaut $11 + \mathfrak{B}(60, \frac{2}{3})$ et F est défini comme $D + 9 + \Delta$ où Δ suit une loi de Poisson $\mathfrak{P}(40)$.


Euclidean
Begin
End
Middle
Y
X

La hauteur du plateau est variable, et suit une loi Gaussienne $\sigma_1\mathfrak{N}(1, 0.2)$ centrée en 1 et d'écart-type ≈ 0.2, avec $\sigma_1 \in -1, 1$. Le pic apparaît en moyenne soit entre les instants 21 et 27 $(11+\mathfrak{B}(15, \frac{2}{3}))$, soit entre les instants 123 et 129 $(139-\mathfrak{B}(15, \frac{2}{3}))$. Sa hauteur varie autour de 1 ou de -1 ($\sigma_2\mathfrak{N}(1, 0.1)$ avec $\sigma_2 \in -1, 1$). A l'issue de cette construction, les instants des séries sont tous perturbés à partir d'un bruit Gaussien $\epsilon(t) = \mathfrak{N}(0, 0.01)$, centré et d'écart-type 0.01. Les séries ainsi construites sont finalement lissées par un filtre moyenneur.

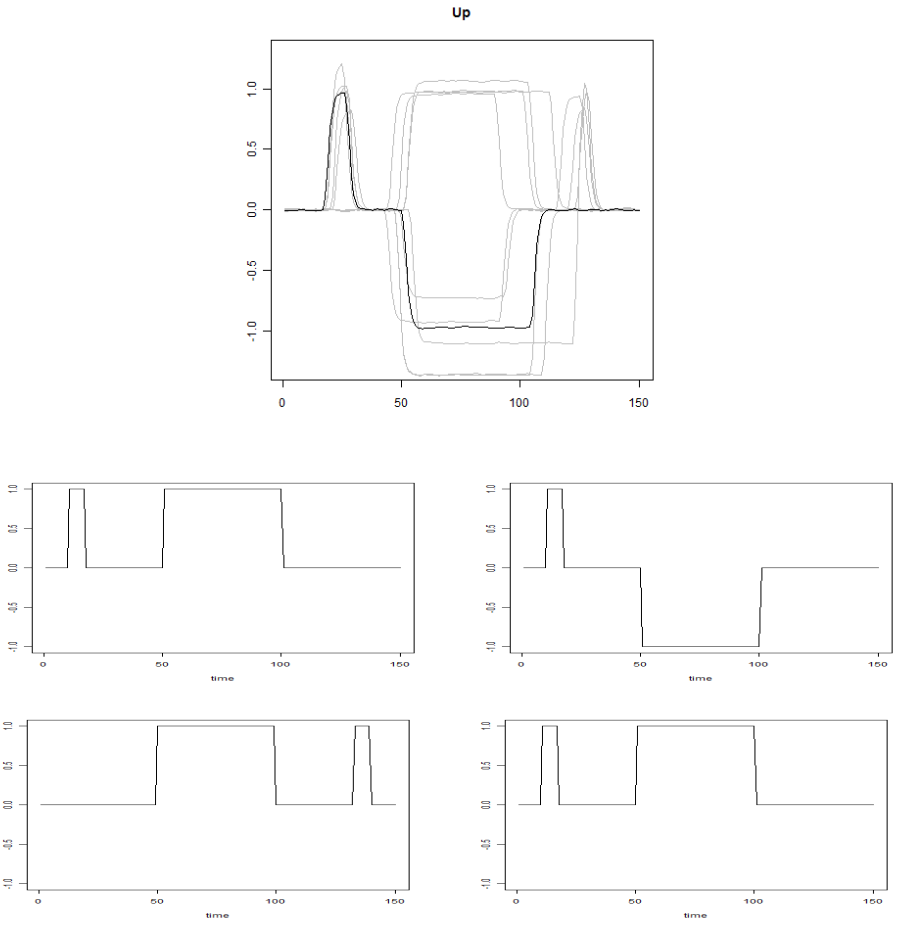
Détail des classes • La classe "Middle" est similaire à la classe Middle du jeu "Begin-Middle-End". Les séries de cette classe ne comporte qu'un plateau central orienté vers le haut.($\sigma_1 = 1$ et $\sigma_2 = 0$ dans le modèle génératif décrit ci-dessus) Ce qui discrimine cette classe des autres est l'absence de pic aux instants correspondant à la réunion des régions des deux pics.

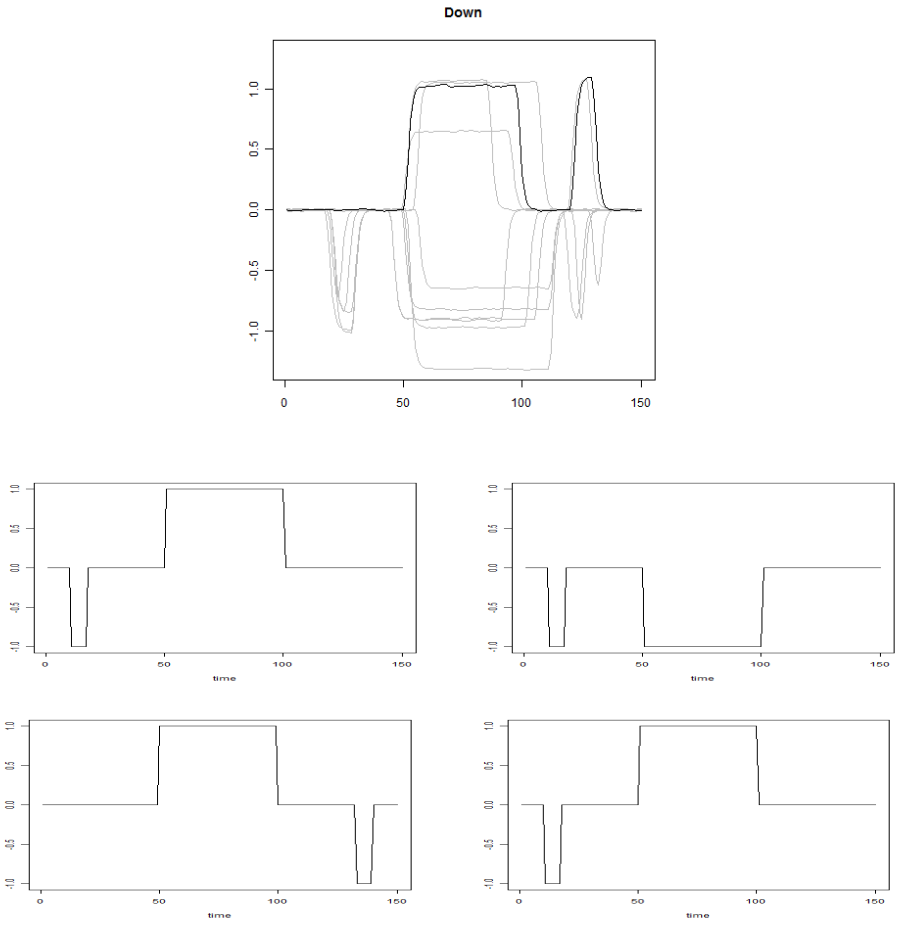


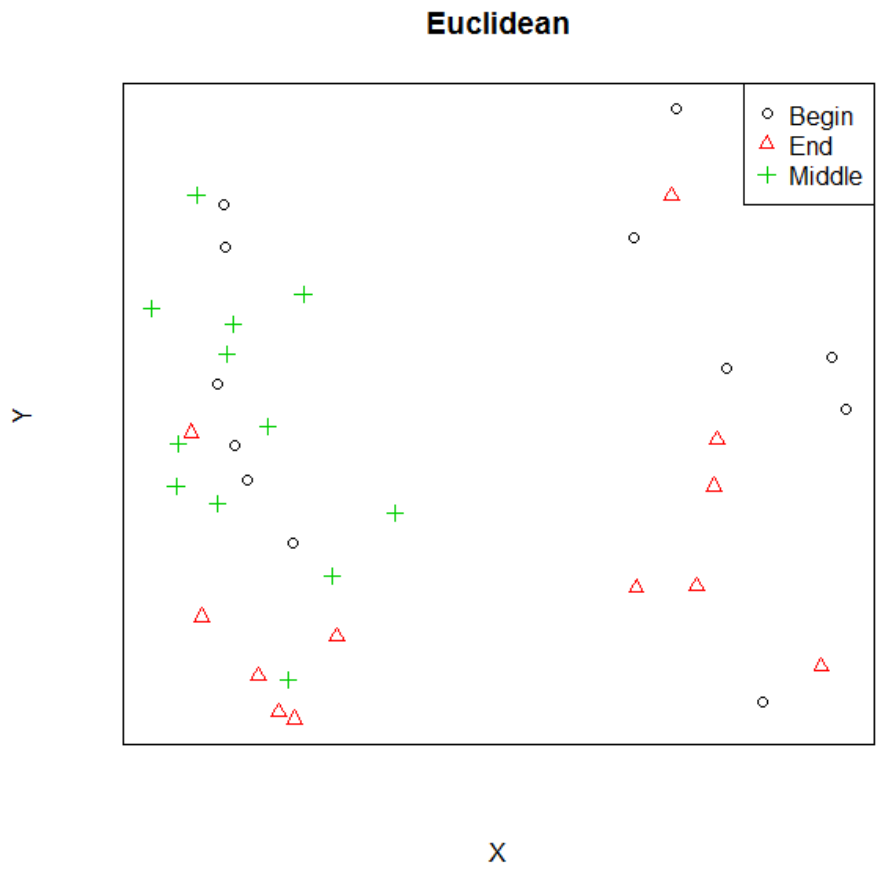
• La classe "Up" est caractérisée par des pics orientés vers le haut et un plateau dont le signe varie ($\sigma_2 \in \{1, -1\}$ avec équiprobabilité et $\sigma_2 = 1$ dans le modèle génératif décrit ci-dessus). Il y a donc quatre types de profils envisageables pour cette classe.

• La classe "Down" est caractérisée par des pics orientés vers le bas et un plateau dont le signe varie ($\sigma_2 \in \{1, -1\}$ avec équiprobabilité et $\sigma_2 = -1$ dans le modèle génératif décrit ci-dessus). Il y a donc quatre types de profils envisageables pour cette classe.

Complexité du jeu de données Comme pour le jeu Begin-Middle-End, la distance euclidienne, calculée entre chaque paire de série, tient compte principalement de la position de la bosse centrale. Nous voyons sur le MDS figure 2 à nouveau deux groupes. Le premier correspond aux séries ayant une bosse centrale vers le haut, le second aux séries ayant une bosse vers le bas. Ce jeu est donc complexe dans le sens où la structure des classes ne s'explique pas sur la lecture des écarts au sein des instants.


Up
1.0
0.5
0.0
-0.5
-1.0
0
50
100
150
time


Down
1.0
0.5
0.0
-0.5
-1.0
0
50
100
150
time


Euclidean
Begin
End
Middle
Y
X

Annexe D

Démonstrations

1 preuves du chapitre 2

Proposition 16 : (Borne de l'indice de Moran)

Si, $\forall i_0, j_0 \in \{1, \ldots, n\} \sum_{i=1}^{n} w_{ij_0} = \sum_{j=1}^{n} w_{i_0 j} = \frac{1}{n}$, alors $I_M \in [-1, 1]$

Preuve

$$|z_i z_j w_{ij}| \leq \frac{1}{2}(z_i^2 w_{ij} + z_j^2 w_{ij}) \tag{101}$$

$$\text{car } (z_i\sqrt{W_{ij}} \pm z_j\sqrt{W_{ij}})^2 > 0 \tag{102}$$

$$\text{D'où } \left|\sum_{j=1}^{n} z_i z_j w_{ij}\right| \leq \sum_{j=1}^{n} |z_i z_j w_{ij}| \leq \frac{1}{2}\sum_{j=1}^{n} z_i^2 w_{ij} + \frac{1}{2}\sum_{j=1}^{n} z_j^2 w_{ij} \tag{103}$$

$$\text{et } \sum_{i=1}^{n}\left|\sum_{j=1}^{n} z_i z_j w_{ij}\right| \leq \sum_{i=1}^{n}\frac{1}{2}z_i^2 \underbrace{\sum_{j=1}^{n} w_{ij}}_{=\frac{1}{n}} + \sum_{i=1}^{n}\frac{1}{2}\sum_{j=1}^{n} z_j^2 w_{ij} \tag{104}$$

$$\left|\sum_{i,j=1}^{n} z_i z_j w_{ij}\right| \leq \sum_{i=1}^{n}\frac{1}{2}z_i^2 + \frac{1}{2}\sum_{j=1}^{n} z_j^2 \underbrace{\sum_{i=1}^{n} w_{ij}}_{=\frac{1}{n}} \tag{105}$$

$$\leq \sum_{j=1}^{n} z_j^2 \tag{106}$$

$$\text{et donc, } -1 \leq I_M \leq 1 \tag{107}$$

•

Proposition 18 :

$$Si\ \forall i_0, j_0 \in \{1,\dots,n\} \sum_{j=1}^{n} w_{i_0 j} = \sum_{i=1}^{n} w_{i j_0} = \frac{1}{n},\ alors\ I_C = \frac{n-1}{n}(1 - I_M)$$

Preuve

$$\begin{aligned}
\sum_{ij} w_{ij}(x_i - x_j)^2 &= \sum_{ij} w_{ij}(x_i - \overline{x} - x_j + \overline{x})^2 \\
&= \sum_{ij} w_{ij} z_i^2 + \sum_{ij} w_{ij} z_j^2 - 2\sum_{ij} w_{ij} z_i z_j \\
&= \sum_{i} z_i^2 \underbrace{\sum_{j} w_{ij}}_{=\frac{1}{n}} + \sum_{j} z_j^2 \underbrace{\sum_{i} w_{ij}}_{=\frac{1}{n}} - 2\sum_{ij} w_{ij} z_i z_j \\
&= 2\sum_{j} z_j^2 - 2\sum_{ij} w_{ij} z_i z_j \\
\text{D'où } \frac{\sum_{ij} w_{ij}(x_i - x_j)^2}{2\sum_j z_j^2} &= 1 - \frac{2\sum_{ij} w_{ij} z_i z_j}{2\sum_j z_j^2} \\
\text{Donc } I_C = \frac{(n-1)\sum_{ij} w_{ij}(x_i - x_j)^2}{2n\sum_j z_j^2} &= \frac{n-1}{n}\left(1 - \frac{2\sum_{ij} w_{ij} z_i z_j}{2\sum_j z_j^2}\right) = \frac{n-1}{n}(1 - I_M)
\end{aligned}$$

Proposition 21 : (Bornes des indices corrigés)
Avec cette modification des indices, l'indice de Geary varie entre 0 et 2, et l'indice de Moran varie entre -1 et 1. Les deux indices sont liés par la relation $I_C* = 1 - I_M*$

Preuve

La preuve des bornes de l'indice de Moran est encore plus directe que la précédente.

$$\begin{aligned}
\left|\sum_{j=1}^{n} z_i z_j w_{ij}\right| &\le \sum_{j=1}^{n} |z_i z_j w_{ij}| \le \frac{1}{2}\sum_{j=1}^{n} z_i^2 w_{ij} + \frac{1}{2}\sum_{j=1}^{n} z_j^2 w_{ij} \\
\sum_{i=1}^{n}\left|\sum_{j=1}^{n} z_i z_j w_{ij}\right| &\le \sum_{i=1}^{n}\frac{1}{2} z_i^2 \sum_{j=1}^{n} w_{ij} + \sum_{j=1}^{n}\frac{1}{2} z_j^2 \sum_{i=1}^{n} w_{ij} \\
\left|\sum_{i,j=1}^{n} z_i z_j w_{ij}\right| &\le \sum_{i=1}^{n} z_i^2 \left(\sum_{j} \frac{w_{ij} + w_{ji}}{2}\right) \\
&\le \sum_{j=1}^{n} N_{jj} z_j^2
\end{aligned}$$

et donc, $-1 \le I_M* \le 1$

Observons la relation qui lie les deux indices.

$$\begin{aligned}\sum_{ij} w_{ij}(x_i - x_j)^2 &= \sum_{ij} w_{ij}(z_i - z_j)^2 \\ &= \sum_{ij} w_{ij} z_i^2 + \sum_{ij} w_{ij} z_j^2 - 2\sum_{ij} w_{ij} z_i z_j \\ &= \sum_i z_i^2 \sum_j (w_{ij}) + \sum_j z_j^2 \sum_i w_{ij} - 2\sum_{ij} w_{ij} z_i z_j \\ &= 2\sum_j z_j^2 \sum_i \frac{w_{ij} + w_{ji}}{2} - 2\sum_{ij} w_{ij} z_i z_j\end{aligned}$$

$$\text{D'où } \frac{\sum_{ij} w_{ij}(x_i - x_j)^2}{2\sum_j N_{jj} z_j^2} = 1 - \frac{2\sum_{ij} w_{ij} z_i z_j}{2\sum_j N_{jj} z_j^2}$$

$$\text{Donc } I_C* = 1 - I_M*$$

•

Proposition 23 : (Indice de Geary corrigé)
L'indice de Geary s'écrit $I_c = \frac{{}^tY(N-\tilde{W})Y}{{}^tYNY}$ *où* $\tilde{W} = \frac{W+{}^tW}{2}$

Preuve :

$$\begin{aligned}{}^tY(N-\tilde{W})Y_{jj'} &= \sum_{i'} y_{i'j} \times \sum_i (N-\tilde{W})_{i'i} y_{ij'} \\ &= \sum_{ii'} \underbrace{N_{i'i} y_{i'j} y_{ij'}}_{=0 \text{ si } i \neq i'} - \sum_{ii'} \tilde{W}_{i'i} y_{i'j} y_{ij'} \\ &= \sum_i N_{ii} y_{ij} y_{ij'} - \sum_{ii'} \frac{W_{i'i} + W_{ii'}}{2} y_{i'j} y_{ij'} \\ &= \sum_i \sum_{i'} \frac{W_{ii'} + W_{i'i}}{2} y_{ij} y_{ij'} - \sum_{ii'} \frac{W_{i'i} + W_{ii'}}{2} y_{i'j} y_{ij'} \\ &= \frac{1}{2}\left(\sum_{ii'} (W_{i'i} + W_{ii'}) y_{i'j} y_{i'j'} - \sum_{ii'} (W_{i'i} + W_{ii'}) y_{i'j} y_{ij'} \right)\end{aligned}$$

$$\begin{aligned}\sum_{ii'} W_{ii'}(x_{ij} - x_{i'j})(x_{ij'} - x_{i'j'}) &= \sum_{ii'} W_{ii'}(y_{ij} - y_{i'j})(y_{ij'} - y_{i'j'}) \\ &= \sum_{ii'} W_{ii'}(y_{ij} y_{ij'} - y_{i'j} y_{ij'} - y_{ij} y_{i'j'} + y_{i'j} y_{i'j'}) \\ &= \sum_{ii'} (W_{ii'} + W_{i'i}) y_{ij} y_{ij'} - (W_{ii'} + W_{i'i}) y_{ij} y_{i'j'}\end{aligned}$$

•

Si P décrit un système de poids, (i.e. $\sum_{i=1}^n p_i = 1$), alors :

$$\sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j (x_i - x_j)^2 = 2\sum_{i=1}^{n} p_i (x_i - \bar{x})^2 \qquad \text{(eq16)}$$

Preuve

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j (x_i - x_j)^2 &= \sum_{i=1}^{n} p_i \sum_{j=1}^{n} p_j (x_i^2 - 2x_i x_j + x_j^2) \\
&= \sum_{i=1}^{n} p_i (x_i^2 \sum_{j=1}^{n} p_j - 2x_i \sum_{j=1}^{n} p_j x_j + \sum_{j=1}^{n} p_j x_j^2) \\
&= \sum_{i=1}^{n} p_i x_i^2 - 2\sum_{i=1}^{n} p_i x_i \bar{x} + \sum_{i=1}^{n} p_i \sum_{j=1}^{n} p_j x_j^2) \\
&= 2\sum_{i=1}^{n} p_i x_i^2 - 2\bar{x}^2 \\
&= 2\sum_{i=1}^{n} p_i (x_i - \bar{x} + \bar{x})^2 - 2\bar{x}^2 \\
&= 2\sum_{i=1}^{n} p_i ((x_i - \bar{x})^2 + \bar{x})^2 + 2(x_i - \bar{x})\bar{x}) - 2\bar{x}^2 \\
&= 2\sum_{i=1}^{n} p_i (x_i - \bar{x})^2
\end{aligned}$$

•

Remarque 24 : (symétrie)
Dans le cas d'une matrice W symétrique, la matrice $N - \tilde{W}$ est la matrice d'une forme quadratique semi-définie positive. La variance locale se présente comme une variance classique limitée aux paires d'observations voisines.

Preuve

$$\begin{aligned}
{}^tY(N - W)Y_{ij} &= \sum_{l} y_{li} \times \sum_{k} (N - W)_{lk} y_{kj} \\
&= \sum_{lk} \underbrace{N_{lk} y_{li} y_{kj}}_{=0 \text{ si } k \neq l} - \sum_{lk} W_{lk} y_{li} y_{kj}
\end{aligned}$$

•

2 preuves du chapitre 3

Proposition 41 : (Calcul de la contribution par la formule des centres mobiles)

$$CM_{ii'}^{ll'} = \sum_{j=1}^{p} \underbrace{\frac{-M_{ii'}^{ll'}}{1-M_{ii'}^{ll'}}(MX_{ij}^{l} + X_{i'j}^{l'})}_{\text{moyenne de } X \text{ et de } MX} \underbrace{(2X_{ij}^{l} - \frac{2-M_{ii'}^{ll'}}{1-M_{ii'}^{ll'}}MX_{ij}^{l})}_{\text{écart entre } X \text{ et } MX}$$

Preuve.

$$\begin{aligned} CM_{ii'}^{ll'} &= \sum_{j=1}^{p}(X_{ij}^{l} - MX_{ij}^{l})^2 - (X_{ij}^{l} - \tilde{MX}_{ij}^{l})^2 \\ &= \sum_{j=1}^{p}(\tilde{MX}_{ij}^{l}[i',l'] - MX_{ij}^{l})(2X_{ij}^{l} - \tilde{MX}_{ij}^{l} - MX_{ij}^{l}) \end{aligned}$$

Or

$$\tilde{MX}_{ij}^{l}[i',l'] = \frac{1}{1-M_{ii'}^{ll'}}(MX_{ij}^{l}) - M_{ii'}^{ll'}X_{i'j}^{l'})$$

Donc

$$\begin{aligned} CM_{ii'}^{ll'} &= \sum_{j=1}^{p}((\frac{1}{1-M_{ii'}^{ll'}} - 1)(MX_{ij}^{l}) - \frac{1}{1-M_{ii'}^{ll'}}M_{ii'}^{ll'}X_{i'j}^{l'})(2X_{ij}^{l} - \tilde{MX}_{ij}^{l} - MX_{ij}^{l}) \\ &= \sum_{j=1}^{p} \underbrace{\frac{-M_{ii'}^{ll'}}{1-M_{ii'}^{ll'}}(MX_{ij}^{l} + X_{i'j}^{l'})}_{\text{moyenne de X et de MX}} \underbrace{(2X_{ij}^{l} - \frac{2-M_{ii'}^{ll'}}{1-M_{ii'}^{ll'}}MX_{ij}^{l})}_{\text{écart entre X et MX}} \end{aligned}$$

■

Proposition 49 :

$$\begin{aligned} \forall\ (i,l) \in \{1,\dots,T\} \times \{1,\dots,n\}\{(i,l,i',l')/CM_{ii'}^{ll'} > 0\} = \varnothing \\ \iff\ (i',l') \in \{1..T\} \times \{1..n\} CM_{ii'}^{ll'} = 0 \end{aligned}$$

Preuve. Le sens $\leftarrow$ est évident, montrons le sens direct par contraposée. Supposons qu'il existe un couple (i',l') tel que $CM_{ii'}^{ll'} \neq 0$. Notons Mx_i^l la moyenne des valeurs prises sur le voisinage de x_i^l.

$$Mx_i^l = \sum_{i_1=1}^{T}\sum_{l_1=1}^{n} m_{ii_1}^{ll_1}x_{i_1}^{l_1}$$

Supposons $Mx_i^l > x$. Alors, $\exists x_{\rangle}^{\updownarrow} = max(x_{i_1}^{l_1} > Mx_i^l)$ (car une moyenne pondérée est toujours majorée par un des termes qui la compose). ■

Proposition 53 :
Le problème de minimisation est un problème convexe.

Preuve. Il est évident que les contraintes définissent un sous-espace convexe de $\mathbb{R}^{n^2T^2}$. Montrons que la fonction $VM(M_{ii'}^{ll'})$ est une fonction convexe de $\mathbb{R}^{n^2T^2}$ dans $\mathbb{R}$. Calculons la matrice Hessienne de VM.

$$H_{(i_1i'_1l_1l'_1)(i_2i'_2l_2l'_2)} = \frac{\partial^2 VM}{\partial M_{i_1i'_1}^{l_1l'_1} \partial M_{i_2i'_2}^{l_2l'_2}} = \quad 0 \text{ si } (i_1, l_1) \neq (i_2, l_2) \tag{108}$$

$$= \quad 2x_{i'_1}^{l'_1} x_{i'_2}^{l'_2} \text{ sinon.} \tag{109}$$

A chaque couple (i, l), on fait correspondre un bloc matriciel $H^{(i_1,l_1,i_2,l_2)} = \left\{ H_{(i_1i'_1l_1l'_1)(i_2i'_2l_2l'_2)}, (i'_1, i'_2) \in 1..T^2, \right.$ dans la matrice Hessienne dont le terme général est : $2x_{i'_1}^{l'_1} x_{i'_2}^{l'_2}$. D'après ce qui précède, les blocs $H^{(i_1,l_1,i_2,l_2)}$ sont nuls si $(i_1, l_1) \neq (i_2, l_2)$. La matrice H est une matrice diagonale par blocs. En posant A la matrice dont la première ligne est constituée du vecteur X multiplié par $\sqrt{2}$, les autres lignes sont nulles, alors les blocs diagonaux sont tous égaux à AA^T. Les blocs diagonaux sont tous des matrices symétriques semi-définies positives. Les blocs non diagonaux étant nuls, la matrice H dans son ensemble est une matrice semi-définie positive. La fonction VM est une fonction convexe.

Ce problème est donc un problème d'optimisation convexe. ■

Proposition 54 :
Le problème de minimisation est un problème discret.

Preuve. L'approche booléenne revient à chercher un sous-ensemble d'arêtes dont le poids associé est directement induit par la taille du voisinage, i.e., par le nombre d'arêtes. Ainsi, le nombre de configurations vérifiant les contraintes est fini. ■

Annexe E

Analyses exploratoires de données contiguës

Nous présentons dans cette annexe des travaux menés en amont de ce travail de thèse sur les méthodes d'analyses exploratoires utilisées dans le cadre de données contiguës. Nous présentons les principales approches. Nous constatons que le choix d'une matrice de voisinage est fondamental dans le cadre de ces approches.

0.1 Rappels sur l'Analyse en Composantes Principales classique

Nous rappellons l'expression du vecteur "moyenne" défini par l'espérance mathématique ainsi que la variance et la covariance d'une variable.

$$\begin{aligned} E(X_j) &= \sum_i m_i x_{ij} \\ Var(X_j) &= E((X_j - E(X_j))^2) = E(Y_j^2) \\ &= \sum_i \left(m_i (x_{ij} - \overline{X_j})^2 \right) \\ Cov(X_j, X_{j'}) &= \sum_i \left(m_i (x_{ij} - \overline{X_j})(x_{ij'} - \overline{X_{j'}}) \right) \end{aligned}$$

Matriciellement, ces résultats s'expriment de la manière suivante :

$$\begin{aligned} E(X) &= \mathbb{1}_n D X \\ Var(X) &= ((I_n - \mathbb{1}_n^t \mathbb{1}_n D)X)^t D((I_n - \mathbb{1}_n^t \mathbb{1} D)X)Q \\ &= Y^t D Y Q \end{aligned}$$

L'ACP revient à rechercher les vecteurs propres associés aux valeurs propres maximales de la matrice de variance-covariance.

Lien avec la physique L'ACP repose sur la correspondance qui existe entre le moment d'inertie et la variance, où le poids statistique d'un individu correspond à la masse physique, et la valeur observée est une distance à un point. Toutefois, le cadre statique ne prend pas en compte les liens existants entre des individus successifs. Le travail sur des séries temporelles

nous amène donc à nous intéresser à un problème plus général de dynamique. L'objectif de ce qui suit est de généraliser notre approche, dans le cadre des séries temporelles, en s'inspirant des grandeurs classiques en cinétique telles que la vitesse.

0.2 Analyses fondées sur les définitions de la variance

0.2.a Approche discriminante fondée sur ces variances

L'analyse discriminante consiste à décomposer la variance en deux parties, la variance entre les classes et la variance au sein des classes. De la même façon, nous séparons les individus entre voisins et non voisins. Nous maximisons ensuite le rapport des variances entre voisins sur la variance totale. Il faut définir la notion de voisins et la notion de non-voisins. Nous pouvons dans le calcul de la variance totale considérer tous les couples de points ; ceci aura pour effet d'éloigner les voisins. Nous pouvons n'en considérer que quelques-uns, ce qui revient à éloigner les points voisins, et à rapprocher les non-voisins.

0.3 Analyses fondées sur une transformation des données

Dans les deux analyses précédentes, un problème particulier se présente. Bien que les matrices soient diagonalisables, les individus du tableau initial X n'apparaissent plus. L'analyse locale est équivalente à une ACP d'un nouveau tableau dont les individus sont des couples d'observations voisines. Ainsi, une alternative a été proposée par Benali et Escofier (1990), consistant à proposer une métrique pour représenter les individus dans un nouvel espace. Les propositions consistent à envisager deux types de métriques, la première lisse les données en fonction du voisinage, tandis que la seconde substitue aux données initiales les différences observées entre la valeur x_i et les valeurs prises au sein de son voisinage.

0.3.a Analyse lissée

L'analyse lissée permet d'analyser les tendances générales des données en éliminant les fluctuations locales. Il s'agit de l'analyse en composantes principales classique de la matrice de données

$$X_{liss} = \qquad\qquad N^{-1} \times W \times X \tag{110}$$

avec N la matrice des poids des individus, W la matrice des poids de voisinage. Cette analyse s'apparente à l'analyse inter-classes. En effet, dans le contexte de l'analyse inter, les classes d'individus sont vues comme des cliques contenant tous les éléments de la classe, et on substitue aux valeurs la moyenne des valeurs observées sur toute la classe.
Notons qu'ici, ce n'est plus une structure en cliques ; les classes sont recouvrantes. On substitue à chaque point la moyenne de ses voisins.
La méthode tient compte de la structure de contiguïté dans le calcul de la moyenne de voisinage. Elle diffère des analyses locales et globales dans le sens où on ne recherche pas une projection optimisant un critère particulier, mais une projection maximisant la variance, dans un nouvel espace (classique au sens de l'ACP), et possèdant toutes les propriétés des composantes principales, en termes d'orthogonalité notamment. Elle recherche la variable pour laquelle la variabilité globale est la plus importante, indépendamment de la similarité locale.

0.3.b Analyse des différences locales

L'analyse des différences locales (Benali et Escofier, 1990) est fondée sur un principe analogue à l'analyse lissée. L'objectif est d'analyser les fluctuations locales en éliminant les variations générales de voisinage. Il s'agit de l'analyse en composantes principales classique de la matrice de données

$$X_{DiffLo} = X - N^{-1} \times W \times X \tag{111}$$

avec N la matrice des poids des individus, W la matrice des poids de voisinage. Cette analyse s'apparente à l'analyse intra-classe. En effet, dans le contexte de l'analyse intra, les classes d'individus sont vues comme des cliques contenant tous les éléments de la classe, et nous étudions les données centrées par la moyenne de voisinage. A nouveau, nous généralisons l'approche intra à des classes recouvrantes. Chaque point est centré par la moyenne de ses voisins.
Comme pour l'analyse lissée, la méthode tient compte de la structure de contiguïté différemment de l'analyse locale, dans le sens où il s'agit d'une véritable analyse en composantes principales.
Cette approche ne recherche pas la variable pour laquelle les voisins sont le plus éloignés, mais la variable pour laquelle la variabilité locale est la plus forte, indépendamment de la structure globale.

0.3.c Analyse de Mom

L'analyse classique, qui consiste à éloigner les centres de gravité des séries et à rapprocher tous les points au sein d'une même classe autour du centre de gravité ne tient pas compte de la structure particulière des séries temporelles au niveau des courbes. Nous aimerions étendre le cadre discriminant en utilisant la notion de contiguïté.
Les travaux de Mom (1988) dans sa thèse généralisent l'analyse discriminante à une structure plus générale. Il définit une métrique qui ne limite plus l'analyse à la recherche de directions compatibles avec une répartition des données en classes, mais plus généralement, il recherche des directions adaptées à la structure de graphe décrivant le voisinage. C'est une généralisation de l'analyse factorielle discriminante, où la répartition en classes est le cas particulier d'un graphe de voisinage constitué de cliques.
Il s'inspire de l'analyse de contiguïté de Lebart et décompose la variance totale $V_T = X' \frac{nI_n - U}{n^2} X$ en fonction de V_L et $V_{L'}$ avec $V_L = \frac{{}^tX[(N-W)]X}{K}$ et $V_{L'} = \frac{{}^tX[(N'-W')]X}{K'}$, où W est la matrice booléenne associée au graphe (ou matrice d'incidence), N la matrice diagonale où le $j^{\text{ième}}$ élément est le nombre de voisins du sommet j, et K le nombre total d'arêtes du graphe. N', W' et K' sont les correspondants pour le graphe complémentaire. Mom obtient alors la relation

$$V_T = \frac{K}{K+K'} V_L + \frac{K'}{K+K'} V'_L$$

qui assure une équivalence entre les vecteurs propres de $\frac{V_L}{V_T}$ et de $\frac{V_L}{V_{L'}}$. Dans sa thèse, Mom propose une métrique adaptée de la forme

$$M = N^{-1}(N-W)'D(N-W)N^{-1}$$

Critique de cette analyse Si le sens donné à la variance intra est assez claire, en ce qu'il s'agit de la matrice des différences locales, le sens donné à la variance inter est moins clair. Une autre critique à évoquer vient du fait que à nouveau, tout repose sur un type de relation de voisinage. On ne distingue pas le lien temporel du lien d'appartenance à une classe. Enfin, l'approche de Mom repose sur la matrice d'adjacence du graphe, et ne tient pas compte de la pondération des arêtes.

Proposition d'une nouvelle approche L'idée qu'il y a derrière l'AFD est celle de rapprocher les individus d'une même classe et de séparer au mieux les différentes classes. L'approche de Mom généralise l'AFD en cherchant à rapprocher les individus voisins et à éloigner les individus non-voisins. Il considère le fait d'être non-voisins comme le fait de ne pas avoir d'arête entre les deux points dans le graphe. Nous proposons une nouvelle relation de voisinage consistant à définir à la fois les voisins, et les non-voisins.
Pour cela, nous développons l'idée de Lebart de décomposer la variance totale en deux variances, mais ici, nous proposons de modifier à la fois la métrique de Lebart pour la variance intra, et la métrique pour la variance totale.
Soit W_1 et W_2 deux matrices d'incidence.
Nous cherchons les axes qui rapprochent les sommets reliés par les arêtes de W_1 et éloignent les sommets reliés dans W_2. Notons V_{L_1} et V_{L_2} respectivement les variances fondées sur l'indice de Geary pour les voisinages W_1 et W_2. Nous cherchons à maximiser le rapport $\frac{V_{L_1}}{V_{L_1}+V_{L_2}}$.
Nous allons pour cela nous assurer du fait que la matrice $V_{L_1} + V_{L_2}$ est inversible.
Posons $\tilde{N} = N_1 + N_2$ et $\tilde{W} = W_1 + W_2$.
Naturellement, $\tilde{N} + \tilde{W} = \tilde{N}(I + \tilde{N}^{-1}\tilde{W})$. Dans le cas où la matrice $\tilde{W}$ est symétrique, la forme quadratique associée à $N - W$ est positive, du fait de la positivité de l'indice de Geary. Cependant, elle n'est pas forcément définie positive. En effet, dans le cas où la matrice de voisinage est constituée de sous-graphe disjoints, un vecteur constant sur chaque sous-graphe annule la forme quadratique. En pratique, un tel cas se produit avec une probabilité nulle. De façon générale, quand n est plus grand que p, la matrice ${}^tY(\tilde{W} + \tilde{N})Y$ est inversible (presque sûrement). Remarquons qu'au contraire de l'analyse discriminante, si la matrice n'est pas inversible, nous ne pouvons pas nous contenter de faire une ACP au préalable, car des données décorrélées au sens de la métrique D ne le seront pas forcément au sens de la métrique $N - W$.

0.4 Résultats

Nous étudions sur deux jeux de données simulées les effets des analyses précédentes. Dans cette étude, nous considérerons dans un premier temps une structure de voisinage temporel (deux instants sont voisins s'ils sont consécutifs). Nous étudierons dans un second temps les effets d'autres structures.

0.4.a Etude des indices de Moran et de Geary

Nous avons dans un premier temps considéré un ensemble de séries définies sur deux variables assez basiques.

Présentation du jeu de données L'objectif de cette partie est d'illustrer sur un jeu de données basiques les forces et les faiblesses des méthodes d'analyse exploratoire de données contiguës pour faire de l'exploration sur des séries temporelles. Nous proposons un jeu de données constitué de trente courbes, exprimées selon deux variables notées X et Y, et réparties en trois classes. X est une variable assez lisse et Y une variable contenant un événement saillant. Pour chaque classe, on définit un individu type, pour lequel l'évolution temporelle sur chaque variable est gaussienne.

$$X(t) = \alpha \times exp(-(t-m)^2/\gamma) \tag{112}$$

où t est le vecteur temps qui évolue entre -1 et 1, α est l'intensité de la variable (α peut être négatif), m décrit la position de la bosse, et gamma sa concavité. Nous construisons les autres séries en introduisant du bruit pour les 3 paramètres de la variable.

De façon générale, nous construisons les deux variables selon le schéma suivant : la première variable est plutôt plate, tandis que la deuxième variable est plus amplifiée et que la bosse est plus étroite. Nous visualisons ces séries sur la figure 0.4.a. Nous observons sur la figure que les trois séries ont été placées dans des domaines de valeurs bien distincts, de sorte qu'il existe une direction pour laquelle les variables se séparent nettement.

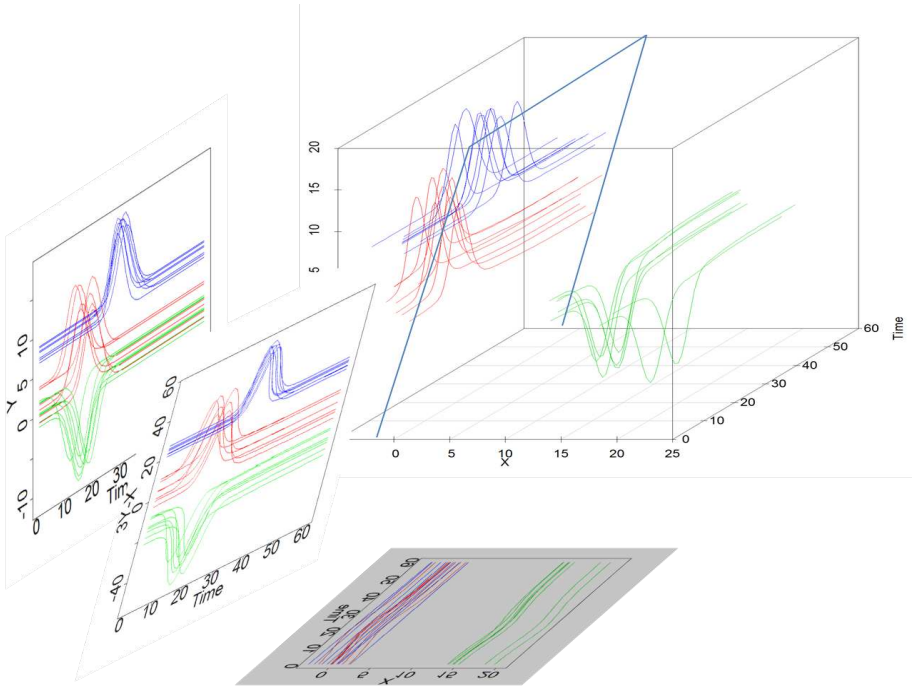


FIGURE 69 – Présentation des séries et axe de séparation

Les classes sont chevauchantes tant pour X que pour Y ; cependant, il existe une direction où les trois classes sont nettement séparées (3Y-X), représentée sur la figure 0.4.a.

Nous observons sur la figure 70 les projections selon les approches globales et locales associées à la structure de voisinage temporel. L'approche locale recherche la variable qui maximise l'indice de Geary. C'est la variable Y. Ceci traduit le fait que l'indice de Geary recherche la variable la plus fluctuante (ici, celle possédant un événement saillant). Au contraire,

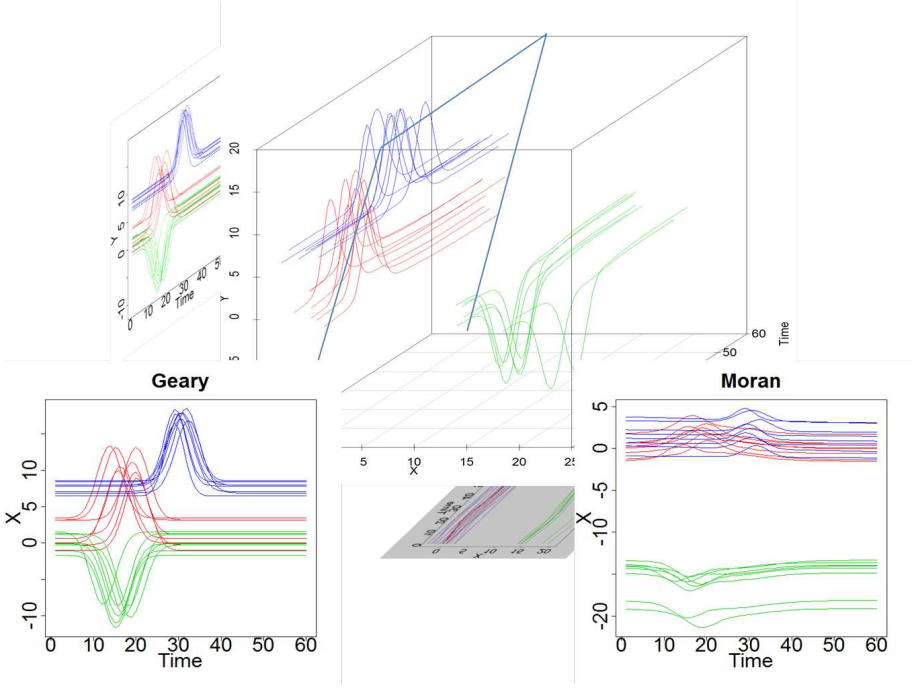


FIGURE 70 – Indices de Moran et de Geary de ces séries

l'approche globale recherche la variable qui maximise l'indice de Moran. C'est la variable X. Ceci traduit le fait que l'indice de Moran recherche la variable la plus lisse.

Lorsqu'on effectue l'analyse des séries temporelles, la méthode fondée sur l'indice de Moran va extraire l'axe de la plus faible amplitude, tandis que la méthode fondée sur l'indice de Geary va extraire au contraire la direction des amplitudes les plus fortes.
La figure 70 visualise le premier vecteur propre de l'analyse de Moran et de celle de Geary, dans le cadre d'une structure de voisinage temporel où les instants voisins sont ceux qui correspondent à des instants consécutifs. Intuitivement, nous attendons d'une bonne approche exploratoire qu'elle mette en avant l'axe qui sépare nettement les trois classes, à l'instar de la projection observée à la figure 0.4.a.
Nous voyons bien sur cet exemple que les deux méthodes n'arrivent pas à trouver l'axe qui sépare au mieux les séries temporelles. Nous remarquons cependant que l'approche locale retrouve l'allure des courbes. C'est un axe qui cherche à réduire au minimum la déformation des courbes, au contraire de l'indice de Moran qui cherche à maximiser la dispersion des courbes.

0.4.b Etude des approches factorielles

Le jeu précédent étant un peu simpliste, nous proposons un jeu simulé plus complexe. Ce jeu est constitué de vingt séries réparties en trois classes et observées selon quatre variables. La figure 71 permet de visualiser ces quatre variables :

- La première variable, notée XI, en référence à l'indice I de Moran, est une variable lisse, i.e., ayant un indice de Moran élevé.
- La seconde variable, notée XC, en référence à l'indice c de Geary, est une variable oscillante, i.e., ayant un indice de Geary élevé.

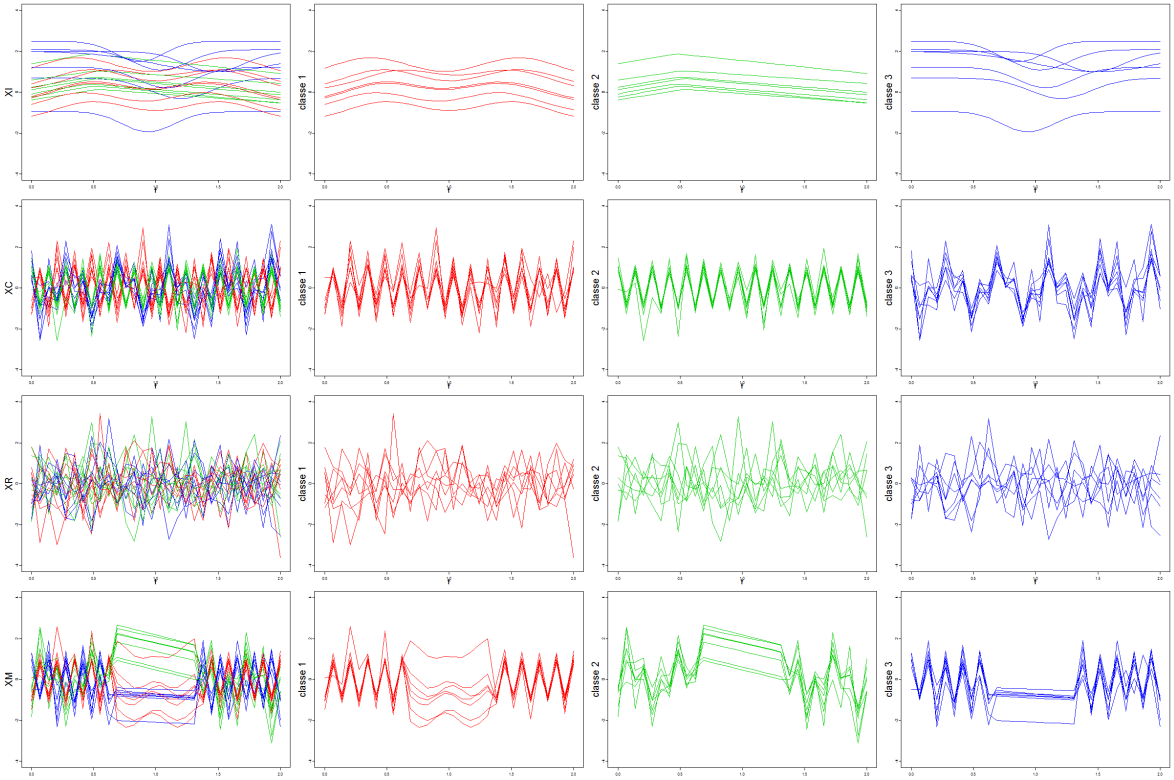
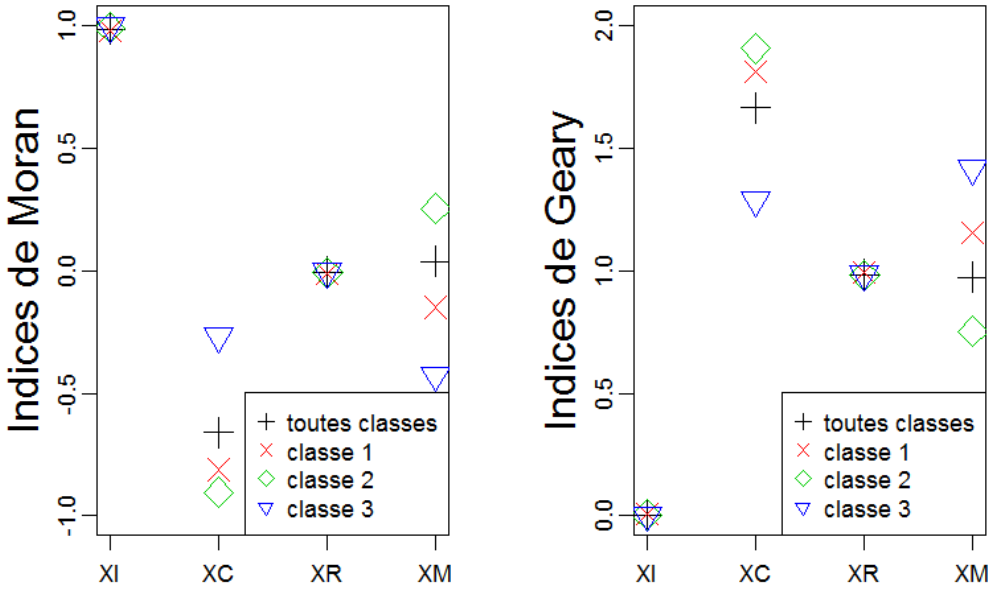


FIGURE 71 – Séries temporelles multivariées simulées

– La troisième variable, notée XR, pour Random, est une variable aléatoire Gaussienne.
– La quatrième variable, notée XM, pour Mixte, est une variable aléatoire composée d'une région lisse située entre deux régions oscillantes.

Nous observons sur la figure 72 la répartition des indices de Moran et de Geary des quatre variables définies ci-dessus. La variable XI a un indice de Moran fort et un indice de Geary faible ; au contraire, la variable XC a un indice de Geary fort et un indice de Moran faible. Ces deux résultats étaient attendus, par construction de ces deux variables. Les indices de Moran et de Geary des deux variables XR et XM sont moyens, comme attendu. Pour XR, c'est dû au fait qu'une variable aléatoire Gaussienne n'est pas liée à une structure de voisinage particulière. Pour la variable XM, les zones où Geary est fort compensent celles où Moran l'est.

La figure 73 donne les premières composantes principales issues des analyses factorielles décrites aux sections 0.2 et 0.3, tandis que la figure 74 affiche l'importance de chacune des variables pour la construction des deux premières composantes principales, à travers le cercle de corrélation, qui indique la corrélation entre une variable et les deux premières composantes principales. Nous remarquons que dans le cercle de corrélation de l'analyse locale, figure une variable sortant du cercle de corrélation. C'est dû au fait que les vecteurs propres ne sont pas nécessairement orthogonaux dans le cadre des analyses locales et globales, au contraire des trois autres approches, qui sont toutes trois des ACP. Nous laissons les cercles de corrélation pour pouvoir comparer les résultats, bien que pour ces deux approches, il faille être attentif à ce point.



(a) Indices de Moran (b) Indices de Geary

FIGURE 72 – Indices spatiaux des variables simulées

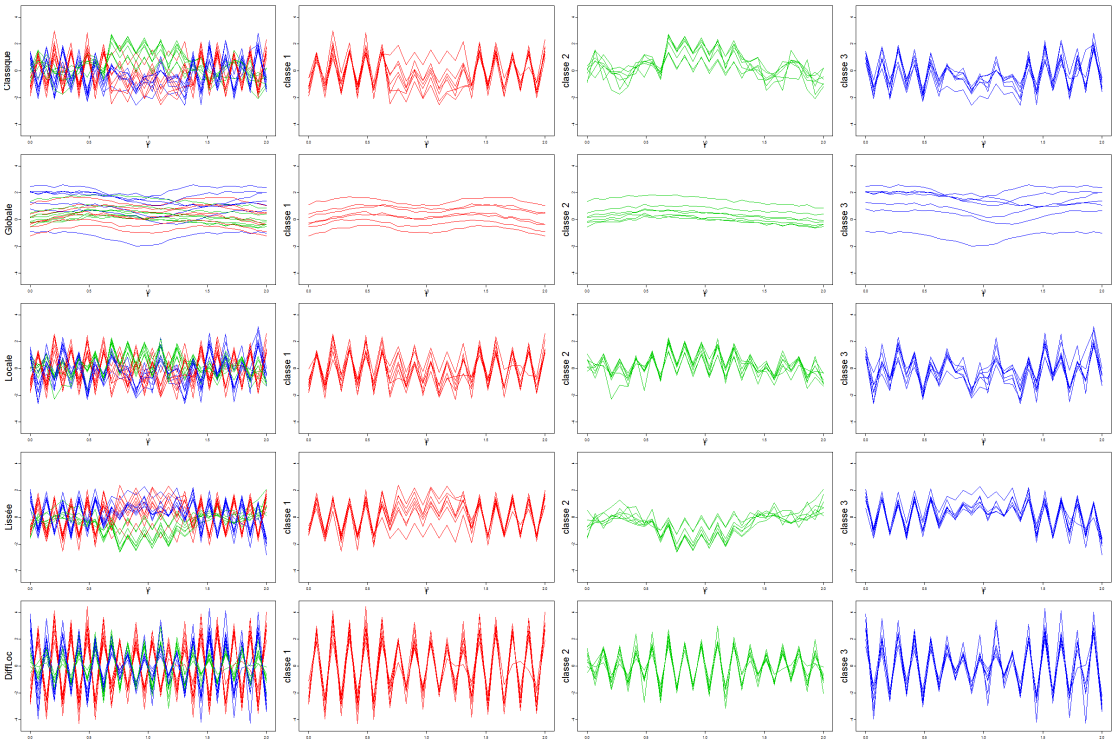


FIGURE 73 – Composantes principales des analyses factorielles

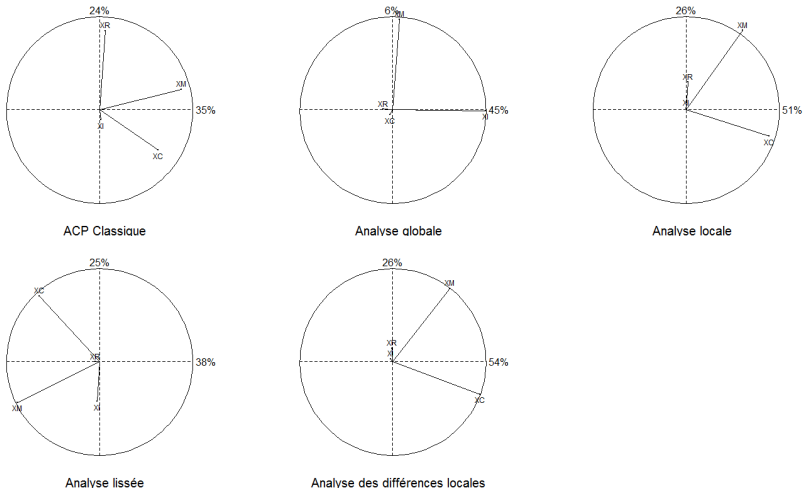


FIGURE 74 – Cercles des corrélations des analyses factorielles

L'approche classique favorise la série mixte, car elle présente de fortes différences entre les instants où apparaissent des oscillations et les instants des plateaux. En effet, l'ACP classique considérant tous les couples d'instants, elle couple les trois régions. En particulier, la partie lisse, en décalage avec l'origine, fait fluctuer l'espérance de la variable. Ainsi, les écarts sont plus importants. La variable mixte est donc celle qui a une variabilité maximale, puis dans une moindre mesure, les variables XR et XC. En effet, pour ces deux variables, la moyenne globale est autour de 0. Les écarts entre les points et le centre sont moyens pour tous les points.
L'analyse globale fait ressortir la variable XI comme celle qui maximise la variance. Ce résultat était attendu car cette variable est celle qui a l'indice de Moran maximal. De manière symétrique, l'analyse locale favorise la variable XC. Les deux approches favorisent dans une moindre mesure la variable XM, du fait de l'existence de régions maximisant l'un ou l'autre des indices.
L'analyse lissée et l'analyse des différences locales font ressortir les variables XC et XM. Ce sont les variables pour lesquelles il y a le plus de différences entre observations voisines. La variable XI fluctue très peu sur son voisinage, et les fluctuations de la variable XR ne sont pas portées par le voisinage, ce qui explique qu'elle n'est pas sélectionnée. Nous remarquons sur la figure 73 que pour ces deux approches, la première composante principale n'est pas une variable initiale. En effet, pour ces deux analyses, les données sont modifiées en amont. Cependant, les méthodes ne permettent pas de séparer les classes de séries. En effet, ceci vient du fait que les analyses de données contiguës considèrent comme objet d'étude les points correspondant aux différents instants des séries, liés entre eux par la relation de voisinage, alors que l'objet de l'étude devrait plutôt être la série.
Les analyses factorielles étudiées précédemment sont des analyses intra-série, alors que nous aimerions faire de l'exploration inter-séries. En considérant que les différents instants d'une même série jouent un rôle similaire, nous pouvons rassembler tous ces points au sein d'une classe et nous avons donc une classe associée à chaque série. Séparer les séries sous cet angle devient un problème d'analyse discriminante.

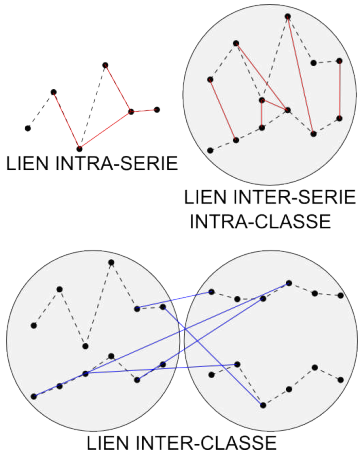


FIGURE 75 – Trois types de liens entre séries

AFD Name	W	Fonction objectif	Type d'approche
AFD1	$W = (O, I, O)$	$max(\frac{V_I(W)}{V_T})$	Variance globale
AFD2	$W1 = (J, I, O)$	$max(\frac{V_I(W1)}{V_I(W1)+V_I(W2)})$	
	$W_2 = (O, O, I)$		
AFD3	$W = (O, O, U)$	$max(\frac{V_c(W)}{V_T})$	Variance locale
AFD4	$W_1 = (I, O, I)$	$max(\frac{V_c(W_1)}{V_c(W_1)+V_c(W_2)})$	
	$W_2 = (O, I, O)$		
AFD5	$W = (O, U, O)$	$max(\frac{VB(W)}{V_T})$	Variance de Mom

TABLE 5 – Approches discriminantes

0.4.c Etude des analyses discriminantes

Nous étudions à présent certaines approches discriminantes. L'approche discriminante dépend pour beaucoup du choix du voisinage. Nous avons sélectionné plusieurs structures particulières de voisinage, que nous résumons dans le tableau 1.

En reprenant les notations introduites à la section 3.1, nous définissons une structure de voisinage par un triplet dont le premier élément décrit la structure d'appariement entre une série et elle-même (lien intra-série), le second élément décrit la structure d'appariement entre deux séries de la même classe (lien inter-séries intra-classe), et le troisième élément décrit la structure d'appariement entre deux séries de classes différentes (lien inter-classes, voir figure 75).

Les figures 76 et 77 donnent la première composante principale des six approches discriminantes présentées dans le tableau 5, ainsi que les cercles de corrélations pour toutes les approches. Nous observons au sein des six analyses en composantes principales présentées, trois types de structure. La première approche, l'analyse factorielle discriminante classique, favorise la variable XM, pour laquelle les valeurs moyennes prises au sein de chaque classe sont les plus séparées. L'approche AFD-5 donne les mêmes résultats ; en effet, la différence entre les deux approches se limite au fait de considérer ou pas la série couplée avec elle-même. L'approche AFD-4 construit la première composante principale majoritairement autour de la variable XC. Le choix du voisinage consiste à coupler les séries instant par instant. La

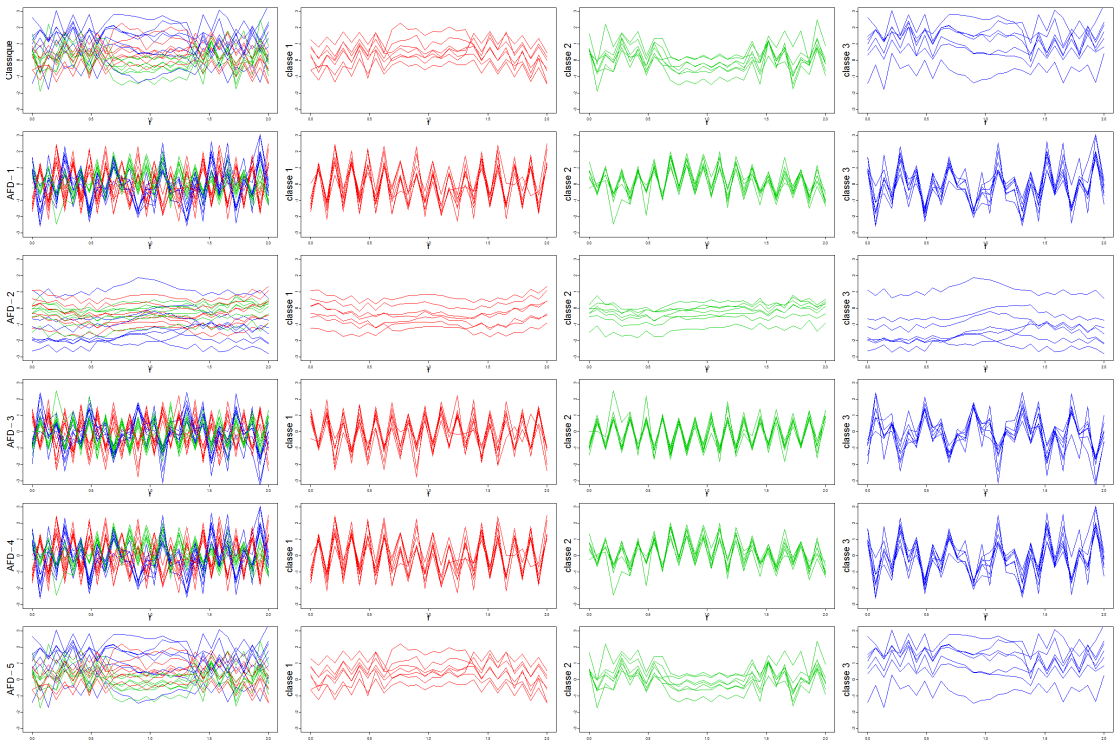


FIGURE 76 – Composantes principales des analyses discriminantes

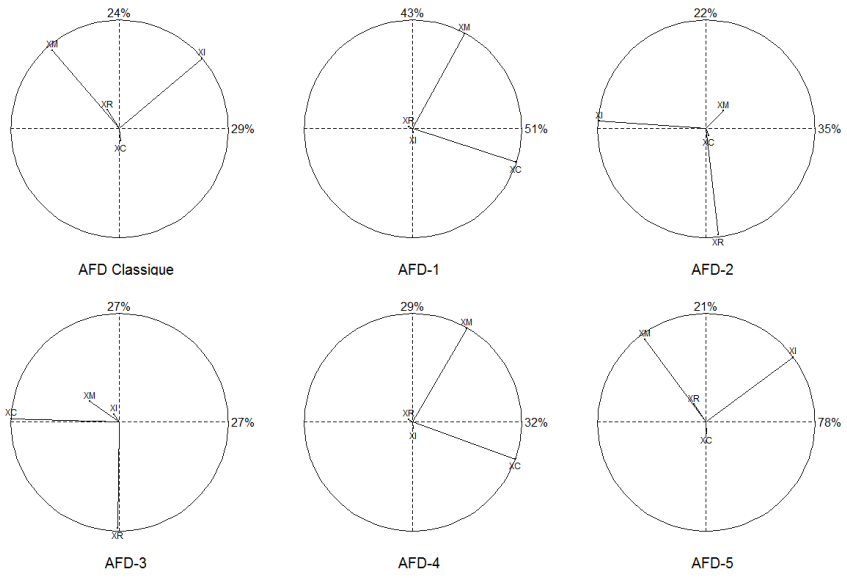


FIGURE 77 – Cercles des corrélations des analyses discriminantes

variable XC est la variable qui rapproche le plus les mêmes instants des séries au sein des classes, et les sépare au mieux entre les classes. Nous remarquons en effet qu'à chaque instant, les classes sont bien séparées pour la variable XC. Le fait que l'approche AFD-1 donne des résultats similaires découle du fait que les deux approches sont complémentaires, l'une étant fondée sur l'approche globale pour un voisinage donné, et la seconde sur l'approche locale pour le voisinage complémentaire.
L'approche AFD-2 construit la première composante principale majoritairement autour de la variable XI. Le choix du voisinage consiste à coupler les séries instant par instant, et à considérer une structure temporelle au cœur des séries. La variable maximisant l'indice de Moran selon cette structure est la variable XI.
L'approche AFD-3 construit la première composante principale majoritairement autour de la variable XC. Le choix du voisinage consiste à coupler les séries selon l'ensemble des instants pour des séries de classes différentes. C'est la variable maximisant l'indice de Geary pour des séries de classes différentes.

Synthèse Les méthodes que nous venons d'étudier reposent toutes sur la définition d'une structure de voisinage pour apparier les paires de séries. Ces approches permettent de mettre en avant des variables ayant certaines propriétés vis-à-vis de la structure de voisinage. Cependant, les appariements utilisés dans ces dernières approches ont été définis de manière ad hoc ; il est nécessaire de définir la structure du voisinage associée à l'ensemble ou à la partition de séries temporelles. La définition de ces appariements est un point important qui a été abordé dans le manuscrit.

Bibliographie

Abdulla, W., Chow, D. et Sin, G. (2003). Cross-words reference template for dtw-based speech recognition systems. *In TENCON 2003. Conference on Convergent Technologies for Asia-Pacific Region*, volume 4, pages 1576–1579. IEEE.

Agrawal, R., Faloutsos, C. et Swami, A. (1993). Efficient similarity search in sequence databases. *Foundations of Data Organization and Algorithms*, pages 69–84.

Banet, T. et Lebart, L. (1984). Local and partial principal component analysis (pca) and correspondence analysis (ca). *IA f. S. Computing., editor. COMPSTAT*, 84:113–123.

Benali, H. et Escofier, B. (1990). Analyse factorielle lissée et analyse factorielle des différences locales. *Revue de statistique appliquée*, 38(2):55–76.

Boyer, L. (2011). *Apprentissage probabiliste de similarités d'édition*. Thèse de doctorat.

Brudno, M., Steinkamp, R. et Morgenstern, B. (2004). The chaos/dialign www server for multiple alignment of genomic sequences. *Nucleic acids research*, 32(suppl 2):W41–W44.

Cai, L., Juedes, D. et Liakhovitch, E. (2000). Evolutionary computation techniques for multiple sequence alignment. *In Evolutionary Computation, 2000. Proceedings of the 2000 Congress on*, volume 2, pages 829–835. IEEE.

Chan, K. et Fu, A. (1999). Efficient time series matching by wavelets. *In Data Engineering, 1999. Proceedings., 15th International Conference on*, pages 126–133. IEEE.

Chouakria-Douzal, A. (2003). Compression technique preserving correlations of a multivariate temporal sequence. *Advances in Intelligent Data Analysis V*, pages 566–577.

Chouakria-Douzal, A. et Nagabhushan, P. (2007). Adaptive dissimilarity index for measuring time series proximity. *Advances in Data Analysis and Classification*, 1(1):5–21.

Cliff, A. et Ord, K. (1972). Testing for spatial autocorrelation among regression residuals. *Geographical Analysis*, 4(3):267–284.

Coleman, R. (1982). *Étude des espaces de Lorentz : leurs relations avec les espaces intersections et unions*. Thèse de doctorat, Université scientifique et médicale de Grenoble.

Cooley, J. et Tukey, J. (1965). An algorithm for the machine calculation of complex fourier series. *Math. Comput*, 19(90):297–301.

Cuturi, M., Vert, J., Birkenes, O. et Matsui, T. (2007). A kernel for time series based on global alignments. *In Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 2, pages II–413. IEEE.

Dayhoff, M. et Orcutt, B. (1979). Methods for identifying proteins by using partial sequences. *Proceedings of the National Academy of Sciences*, 76(5):2170.

Demongeot, J., Drouet, E., Elena, A., Moreira, A., Rechoum, Y. et Sené, S. (2009). Micro-rnas : viral genome and robustness of gene expression in the host. *Philosophical Transactions of the Royal Society A : Mathematical, Physical and Engineering Sciences*, 367(1908):4941–4965.

Devroye, L., Györfi, L. et Lugosi, G. (1996). *A probabilistic theory of pattern recognition*, volume 31. Springer Verlag.

Diallo, A. (2010). *Classification de profils d'expression de gènes : application à l'étude de la régulation du cycle cellulaire chez les eucaryotes.* Thèse de doctorat, Université de Grenoble.

Douzal-Chouakria, A. et Amblard, C. (2011). Classification trees for time series. *Pattern Recognition*.

Dror, O., Benyamini, H., Nussinov, R. et Wolfson, H. (2003). Mass : multiple structural alignment by secondary structures. *Bioinformatics*, 19(suppl 1):i95–i104.

Duret, L. et Abdeddaim, S. (2000). Multiple alignment for structural, functional, or phylogenetic analyses of homologous sequences. *Bioinformatics Sequence structure and databanks*.

Feng, D. et Doolittle, R. (1987). Progressive sequence alignment as a prerequisitetto correct phylogenetic trees. *Journal of molecular evolution*, 25(4):351–360.

Fix, E. et Hodges, J. (1951). Discriminatory analysis, nonparametric regression : consistency properties. Rapport technique, Technical report, USAF School of Aviation Medicine.

Fréchet, M. (1906). Sur quelques points du calcul fonctionnel. *Rendiconti del Circolo Matematico di Palermo (1884-1940)*, 22(1):1–72.

Gaffney, S. et Smyth, P. (2005). Joint probabilistic curve clustering and alignment. *Advances in neural information processing systems*, 17:473–480.

Geary, R. (1954). The contiguity ratio and statistical mapping. *The Incorporated Statistician*, Vol. 5, No. 3 (Nov., 1954):115–127+129–146.

Goutte, C., Toft, P., Rostrup, E., Nielsen, F. et Hansen, L. (1999). On clustering fmri time series. *NeuroImage*, 9(3):298–310.

Hamming, R. (1950). Error detecting and error correcting codes. *Bell System technical journal*, 29(2):147–160.

Hastie, T. et Tibshirani, R. (1996). Discriminant adaptive nearest neighbor classification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 18(6):607–616.

Haussler, D. (1999). Convolution kernels on discrete structures. Rapport technique, Technical report, UC Santa Cruz.

Hautamaki, V., Nykanen, P. et Franti, P. (2008). Time-series clustering by approximate prototypes. *In Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE.

Henikoff, S. et Henikoff, J. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915.

Hogeweg, P. et Hesper, B. (1984). The alignment of sets of sequences and the construction of phyletic trees : an integrated method. *Journal of molecular evolution*, 20(2):175–186.

Hourai, Y., Akutsu, T. et Akiyama, Y. (2004). Optimizing substitution matrices by separating score distributions. *Bioinformatics*, 20(6):863–873.

Hu, J., Ray, B. et Han, L. (2006). An interweaved hmm/dtw approach to robust time series clustering. *In Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 3, pages 145–148. IEEE.

Itakura, F. (1975). Minimum prediction residual principle applied to speech recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 23(1):67–72.

Kaufman, L., Rousseeuw, P. *et al.* (1990). *Finding groups in data : an introduction to cluster analysis*, volume 39. Wiley Online Library.

Keogh, E. et Pazzani, M. (2001). Derivative dynamic time warping. *In First SIAM international conference on data mining*, pages 5–7.

Lebart, L. (1969). Analyse statistique de la contiguïté.

Leibowitz, N., Nussinov, R. et Wolfson, H. (2001). Musta-a general, efficient, automated method for multiple structure alignment and detection of common motifs : application to proteins. *Journal of computational biology*, 8(2):93–121.

Levenshtein, V. (1965). Binary codes capable of correcting spurious insertions and deletions of ones. *Problems of Information transmission*, 1(1):8–17.

Li, C. et Biswas, G. (1999). Temporal pattern generation using hidden markov model based unsupervised classification. *Advances in Intelligent Data Analysis*, pages 245–256.

Listgarten, J. (2007). *Analysis of sibling time series data : alignment and difference detection.* Thèse de doctorat, University of Toronto.

Listgarten, J., Neal, R., Roweis, S., Puckrin, R. et Cutler, S. (2007). Bayesian detection of infrequent differences in sets of time series with shared structure. *Advances in neural information processing systems*, 19:905.

Mahalanobis, P. (1936). On the generalized distance in statistics. *In Proceedings of the National Institute of Sciences of India*, volume 2, pages 49–55. New Delhi.

Mahé, P. et Vert, J. (2009). Graph kernels based on tree patterns for molecules. *Machine learning*, 75(1):3–35.

Mom, A. (1988). *Méthodologie statistique de classification dans les réseaux de transport.* Thèse de doctorat.

Moran, P. (1950). A test for the serial independence of residuals. *Biometrika*, 37(1/2):178–181.

Morgenstern, B., Dress, A. et Werner, T. (1996). Multiple dna and protein sequence alignment based on segment-to-segment comparison. *Proceedings of the National Academy of Sciences*, 93(22):12098.

Needleman, S., Wunsch, C. *et al.* (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453.

Notredame, C. (2002). Recent progress in multiple sequence alignment : a survey. *Pharmacogenomics*, 3(1):131–144.

Notredame, C., Higgins, D., Heringa, J. *et al.* (2000). T-coffee : A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*, 302(1):205–218.

Notredame, C., Holm, L. et Higgins, D. (1998). Coffee : an objective function for multiple sequence alignments. *Bioinformatics*, 14(5):407–422.

Oates, T., Firoiu, L. et Cohen, P. (1999). Clustering time series with hidden markov models and dynamic time warping. *In Proceedings of the IJCAI-99 workshop on neural, symbolic and reinforcement learning methods for sequence learning*, pages 17–21. Citeseer.

Oates, T., Firoiu, L. et Cohen, P. (2001). Using dynamic time warping to bootstrap hmm-based clustering of time series. *Sequence Learning*, pages 35–52.

Owsley, L., Atlas, L. et Bernard, G. (1997). Self-organizing feature maps and hidden markov models for machine-tool monitoring. *Signal Processing, IEEE Transactions on*, 45(11):2787–2798.

Petitjean, F., Ketterlin, A. et Gançarski, P. (2011). A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, 44(3):678–693.

Qamar, A., Gaussier, E., Chevallet, J. et Lim, J. (2008). Similarity learning for nearest neighbor classification. *In Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 983–988. IEEE.

Rabiner, L., Rosenberg, A. et Levinson, S. (1978). Considerations in dynamic time warping algorithms for discrete word recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 26(6):575–582.

Ramsay, J. et Li, X. (1998). Curve registration. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 60(2):351–363.

Ravinder, K. (2010). Comparison of hmm and dtw for isolated word recognition system of punjabi language. *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 244–252.

Rüping, S. (2001). Svm kernels for time series analysis. *In LLWA*, pages 43–50. Citeseer.

Sakoe, H. et Chiba, S. (1971). A dynamic programming approach to continuous speech recognition. *In Proceedings of the Seventh International Congress on Acoustics*, volume 3, pages 65–69.

Sakoe, H. et Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 26(1): 43–49.

Sankoff, D. et Kruskal, J. (1983). Time warps, string edits, and macromolecules : the theory and practice of sequence comparison. *Reading : Addison-Wesley Publication, 1983, edited by Sankoff, David ; Kruskal, Joseph B.*, 1.

Shatsky, M., Nussinov, R. et Wolfson, H. (2002). Multiprot—a multiple protein structural alignment algorithm. *Algorithms in Bioinformatics*, pages 235–250.

Smith, T., Waterman, M. et Fitch, W. (1981). Comparative biosequence metrics. *Journal of Molecular Evolution*, 18(1):38–46.

Srisai, D. et Ratanamahatana, C. (2009). Efficient time series classification under template matching using time warping alignment. *In Computer Sciences and Convergence Information Technology, 2009. ICCIT'09. Fourth International Conference on*, pages 685–690. IEEE.

Thioulouse, J., Chessel, D. et Champely, S. (1995). Multivariate analysis of spatial patterns : a unified approach to local and global structures. *Environmental and Ecological Statistics*, 2:1–14. 10.1007/BF00452928.

Tomasi, G., van den Berg, F. et Andersson, C. (2004). Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *Journal of Chemometrics*, 18(5):231–241.

Vert, J., Saigo, H. et Akutsu, T. (2004). Local alignment kernels for biological sequences. *Kernel methods in computational biology*, pages 131–154.

Warren Liao, T. (2005). Clustering of time series data—a survey. *Pattern Recognition*, 38(11):1857–1874.

Wartenberg, D. (1985). Multivariate spatial correlation : A method for exploratory geographical analysis. *Geographical Analysis*, 17(4):263–283.

Weinberger, K., Blitzer, J. et Saul, L. (2006). Distance metric learning for large margin nearest neighbor classification. *In In NIPS*. Citeseer.

Wilpon, J. et Rabiner, L. (1985). A modified k-means clustering algorithm for use in isolated work recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 33(3):587 – 594.

Xing, E., Ng, A., Jordan, M. et Russell, S. (2002). Distance metric learning, with application to clustering with side-information. *Advances in neural information processing systems*, 15:505–512.

Yang, K. et Shahabi, C. (2004). A pca-based similarity measure for multivariate time series. *In Proceedings of the 2nd ACM international workshop on Multimedia databases*, pages 65–74. ACM.

Yang, K. et Shahabi, C. (2005). A pca-based kernel for kernel pca on multivariate time series. *In Proceedings of ICDM'05 Workshops*, pages 149–156.